Exploring Diatoms Functional and Taxonomic Diversity on a Global Scale Through an Integrative Approach

Thesis

Stazione
Zoologica
Anton Dohrn
Napoli

Exploring diatoms functional and taxonomic diversity on a global scale through an integrative approach

---

Greta Busseni

Thesis submitted for the degree of

Doctor of Philosophy

in Life, Health and Chemical Sciences

September 2018

Open University

ARC Department of Integrative Marine Ecology

ARC Institute Stazione Zoologica Anton Dohrn

Discipline : School of Life, Health and Chemical Sciences

Documentation

Exploring diatoms functional and taxonomic diversity on a global scale through an integrative approach

Greta Busseni

1. Reviewer Colin Brownlee

Marine Biological Association of the UK

University of Southampton

2. Reviewer Maria Luisa Chiusano

Department of 'Agraria'

University 'Federico II' of Naples

Supervisors Daniele Iudicone, Remo Sanges and Chris Bowler

September 2018

Greta Busseni

Exploring diatoms functional and

taxonomic diversity on a global scale

through an integrative approach

Documentation, September 2018

Reviewers: Colin Brownlee and Maria Luisa Chiusano

Supervisors: Daniele Iudicone, Remo Sanges and Chris Bowler

Open University

Discipline : *School of Life, Health and Chemical Sciences*

ARC Institute Stazione Zoologica Anton Dohrn

ARC Department of Integrative Marine Ecology

Walton Hall

MK7 6AA and Milton Keynes

Abstract

Diatoms are a fundamental component of the oceanic ecosystem. Because of the massive primary production they are responsible for they play a pivotal role in several biogeochemical cycles as well as in the marine food web. This high relevance is due to their global distribution together with their seasonal dominance in many of the planktonic communities. This ecological 'success' is granted by their diversity, being diatoms the most diverse microalgae taxa. Notwithstanding the relevant role of this taxa, little is known about their biology and ecology, given the lack of large scale observations and the large number of uncultured species. Here I describe global scale diatom diversity exploring in parallel their taxonomic and their functional diversity. Different methodological frameworks were developed to measure both classes of diversity from meta-omic data. As results, herein it is described the first assessment of diatom taxonomic richness at a global scale together with its statistical modeling using a machine-learning approach. Moreover, a completely new approach to characterize the functional diversity is provided. It is based on the phylogenies of nitrogen transporter marker gene families that proved to be optimal markers of functional traits such as size and resource utilization traits. Finally, the outcome of a numerical modeling exercise was compared to omic taxonomic data with the aim of improving the diatom model types. The whole work has been developed exploiting the unprecedented amount of data provided by the *Tara* Oceans expedition.

Acknowledgement

I would like to thank the Open University (OU) and Stazione Zoologica Anton Dohrn for giving me the opportunity and fellowship to pursue my Ph.D.

I want to thank Dr. Daniele Iudicone, my director of studies, for the support, the teaching and for the guidance through the magnificent complexity of the global Ocean.

I would like to express a special appreciation for my supervisors Dr. Remo Sanges and Dr. Chris Bowler for the guidance and the encouragement given to me during my Ph.D. I would like to express my gratitude to Dr. Luigi Caputi for his daily patience and discussions and to Dr. Maurizio Ribera d'Alcalà for his advices and assistance.

I would like to thanks Dr. Fabio Rocha Jimenez Vieira and Dr. Eric Pelletier for being a reference point during these years with all the encountered technical challenges.I want to thank Dr. Roberta Piredda and Dr. Eleonora Scalco for helping me with all kinds of diatom measures, from the metabarcode to the microscopy ones, and also Dr. Luigi Maiorano who introduced me to the fantastic world of machine learning. I would like to thank Dr. Stephanie Dutkiewicz as well, who let me see the ocean through her numerical model.

I would like to thank the *Tara* Oceans Consortium for giving me access to this incredible data and thus allowing the birth of this project.

I have to thank my mother for being my person, my strength and my role model, my family, Giulia Alessandro and Andrea, for being able to support me each one in your own unique way, and my nephews which in these three years could not make us happier.

I would like to include in these acknowledgments my friends: my musketeers Angela and Mass, because you made this experience fun even in the hardest moments, my SZN friends Arianna, Laura, Ennio, Kata, and all the Saletta people for the countless laughs and lovely comfort you gave me, and my northern allies, Eli, Vale and Arianna, for sharing everything with me notwithstanding the kilometers.

Finally, I want to thank Vincenzo for supporting and standing by me through all this incredible adventure.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Diatom diversity

> *[...] diatoms are*
> *in many respects*
> *the angiosperms of the sea*
>
> — **Parks et al.**
> 2018

The Global Ocean is the largest ecosystem on Earth and the planktonic community living within shows very high diversity, and an estimated concentration of 10^4 to 10^6 organisms in one milliliter of water (Whitman et al., 1998). The rise of a modern eukaryotic phytoplankton community began in the Middle Triassic (Falkowski et al., 2004) and it attained the current ecological and phylogenetic structure only about 65 million years ago (Ma), after a bolide-impact removed a major portion of phytoplankton diversity (Katz et al., 2007). The plankton is often seen as a complex adaptive system strictly linked to the abiotic component and presenting nonlinear interactions across a vast range of space-time scales (Bonachela et al., 2016). While interacting, plankton communities drive and respond to environmental changes, including longer term alterations due to climate change as shifts in temperature, seawater chemistry and variations in stratification and currents (Hays et al., 2005; Doney et al., 2012; Sunagawa et al., 2015).

One leading actor within phytoplankton communities is the diatom group (Bacillariophycaee): unicellular phototrophic eukaryotic micro-organisms living in marine and freshwater environments as well as in soil, snow and even ice. Among the impressive characteristics of this group is that they are responsible for roughly one-fifth of the photosynthesis on Earth (Nelson et al., 1995), and thus also of 20-25% of the oxygen released by photosynthetic organisms, and this alone is a good reason for the large research effort made on this taxa, but reading this introduction you will walk through many other significant aspects of diatoms from their key role in the major ocean biogeochemical cycles to their participation in maintaining the ecosystem stability.

In the literature they are frequently referred to as one of the most ecologically successful groups among photosynthetic eukaryotic micro-organisms (Vanormelingen et al., 2008). They gained this attribute because of their ubiquitous distribution and their dominance in most of the aquatic habitats, two ecological conditions that are usually mutually exclusive. But how can they survive in this variety of habitats, managing both planktonic and benthic lifestyles in almost all aquatic environments, even including temporarily wet terrestrial habitats, and be good competitors in most of these environments? The answer lies in their diversity, with over ca 10^5 species (Mann and Droop, 1996), diatoms are the most diverse unicellular microalgae, at least with our present capability of taxonomic resolution. From the evolutionary point of view, diatoms originated in the early Mesozoic (Kooistra and Medlin, 1996), and then, according to fossil record, they increased their diversity through all the Cenozoic (Gersonde and Harwood, 1990; Sims et al., 2006). Different phylogenies have been proposed through time, the most accepted right now is the one proposed by Medlin and Kaczmarska, 2004 which classify diatoms in three main classes: the most ancestral, the radial centrics Coscinodiscophyceae, followed by the polar centric Mediophyceae and finally by the most recent pennates (Bacillariophyceae). Their diversity is reflected also in their morphology through the frustule structure, i.e., the external porous outer silica

coating the cell, whose function is not only to maintain cell structure but also to act as a protective barrier (Hamm et al., 2003; Mitchell, 2018). Precisely on the shape of this shield diatoms are classified in two main orders: the radially symmetrical centrics (Centrobacillariophyceae; Centrales) and the typical bilateral symmetric pennate (Pennatibacillariophyceae; Pennales) (Round et al., 1990). Nevertheless, diatoms boast an extraordinary morphological diversity which is not limited to their structure, the possibility to own spines or setae, but also to their ability to live in differentially shaped colonies (e.g., flat, spiral, ribbons, fans or star shaped). Another ecologically important aspect of their diversity is their size: marine diatoms have a cell volume spanning almost 9 orders of magnitude with the larger cells exceeding 10^9 μm^3 (Litchman et al., 2009).

Diatoms reproduce via asexual reproduction: this is characterized by a mitosis happening within their cell wall followed by the creation of two smaller valves. As they own two valves, the epitheca (the larger) and hypotheca (the smaller), which are typically overlapped like the lid of a box, the creation of two new valves define the separation of the two daughters cells. The daughters' valves will be the two hypothecas, while the maternal valves will function as epitheca. This reproduction process lead to the shrinking of cells dimensions in the population and this decrease in mean size of daughter cells is called diminution. Consequently, when daughter cells reach a fixed minimum size, the mitosis process stops. To bring back the balance of mean size to bigger sizes, below a certain species-specific dimension threshold cells can either reproduce sexually, forming maximum sized daughters (Chepurnov et al., 2004) or use alternative solutions to restore the maximal size. However, for some of them the life cycle is more complex than this: they not only have the described sexual and asexual cycle but they are also able to form resting stages in the form of spores and resting cells (McQuoid and Hobson, 1996). These spores have thick silica frustules that strongly reduce the buoyancy ability of the species and bring these cells below the euphotic zone down to

the sediment. They can survive for decades in this stage, possibly waiting for optimal condition to germinate, in particular via mixing water processes that bring them again in the euphotic zone if herein they found optimal conditions in terms of N concentration (Kuwata and Takahashi, 1999), temperature and day length (Eilertsen et al., 1995; McQuoid and Hobson, 1996) and light (Shikata et al., 2011). Physiologically, diatoms have wide species-specific differences. Pennate and centric diatoms for example differ not only in their sexual reproduction (Armbrust, 2009) but also in their response to nutrient availability (Strzepek and Harrison, 2004). Moreover, recently the metabolic activity of these algae has been assessed as flexible, finding evidences of mixotrophy in particular species (Villanova et al., 2017).

Diatom community is strongly shaped by the ecological succession process. This latter is discontinuous, either self-controlled or controlled by environmental variability (Odum, 1969) or both. In variable systems such as coastal areas, plankton communities usually change their structure over short time periods like seasons. Species are replaced in time due to biotic or abiotic forcings until the system goes back to the initial state through exogenous forcing (Margalef, 1963). Phytoplankton succession in general begin with strong vertical mixing, leading to a nutrient enrichment of the photic layer. The community evolves then by replacements that largely follows the depletion of various limiting nutrients (Sommer, 1989). These annuals nutrient enrichments lead to typical seasonal blooms. According to Margalef (1978) marine phytoplankton can be classified according to their responses to nutrient availability and turbulence. Diatoms occupy the high nutrient high turbulence conditions, coccolithophorids have intermediate nutrient and turbulence while dinoflagellates dominate in low turbulence and high or low nutrient conditions. Consequently, generally fast-growing diatoms are responsible for the spring bloom in mid-latitudinal regions, while more variety is observed in the series of summer blooms which see as main characters not only diatoms but also flagellates and dinoflagellates, to conclude with the autumn bloom dominated

by diatoms or dinoflagellates (Mallin et al., 1991). Excluding the transient phase of blooms diatom species are adapted to survive also in sub-optimal environmental conditions. Compared to other phytoplanktonic groups diatoms are generally characterized by higher maximum uptake rates of nutrients (Dutkiewicz et al., 2015). Their competitiveness is thus not only given by their relatively fast growth rates, but also by their ability in nutrient uptake and assimilation, allowing them to be the first to take advantage of sudden nutrient supplies (Litchman, 2007). Even if this is the general strategy of diatoms, their size make a strong difference in these parameters. Larger diatoms indeed have slower growth rates, but they still display high maximum nutrient uptake rates. These conditions make large diatoms 'storage adapted' (Sommer, 1984): the size of their vacuole is disproportionally large compared to the relative size of the same structure in small diatoms (Sicko-goad et al., 1984). These vacuoles can store nutrients, in particular nitrate resources, allowing diatoms a nutrient backup in case of starvation conditions until the availability of new supplies (Syrett, 1981; Raven, 1987; Grover, 1991; Stolte and Riegman, 1996).

Diatom responses to the environment are even finer than this. Recent studies have shown their ability to modulate cell activity and colonies structuring (e.g., chain size) as response to grazing activity (Tammilehto et al., 2015; Amato et al., 2018) or turbulence intensity (Iudicone et al., 2016; Amato et al., 2017; Dell'aquila et al., 2017). Around 30 species of diatoms have been identified as harmful to either fisheries, wildlife or people in health or economic terms (Fryxell and Villac, 1999). They can damage producing toxins or exudates or harming physically other organisms due to their cell morphology or high biomass accumulation. Some species are also able to produce toxic secondary metabolites. The function of these toxins is still debated but their production has been found to be likely regulated by the presence of the predator itself (Tammilehto et al., 2015). The first diatom genus discovered producing toxic compounds was *Pseudo-nitzschia*. This toxin,

domoic acid (DA), is naturally occurring and the syndrome of DA poisoning is called Amnesic Shellfish Poisoning as one of the symptoms is amnesia.

From an ecologic point of view, diatoms have evolved into benthic and planktonic forms. While most planktonic diatoms live in the water column, drifting with the water movements, they also have specific adaptations to control their buoyancy (e.g., spines, frustule lightening, and density regulation) (Smetacek, 1985). The greater number of species live a benthic life, spending their existence as epilithic on sand and stones or epiphytic on the vegetation (Kooistra et al., 2007). Planktonic forms of diatoms mostly belong to the centric lineage, but also some major pennate genera, such as *Pseudo-nitzschia*, *Fragilariopsis* and *Thalassiothrix*, adopted this life strategy.

Finally, all this interest in studying diatoms is justified by their involvement in a range of marine ecosystem services. They are at the base of the most productive fisheries, being used as fisheries nourishment. They can be dangerous because of their toxicity and deepening their knowledge is useful for men from both the health and economic point of view. They are main players in biogeochemical cycles as they sequester carbon: which make them fundamental to predict future climate change. And lastly, given the dramatic problems of microplastic pollution in the waters they play once again a role favouring the sedimentation of these pollutants through biofouling activities (Kaiser et al., 2017).

We still don't have a clear understanding of diatom biology and ecology but it is most certainly in the interest of our society to study them, from an economic, health and safety point of view.

1.1.1 Ecological definition of diversity

Nowadays biological diversity, also expressed as biodiversity, can be found in the dictionary defined as 'the variation of life forms within a given ecosystem'. While the idea behind this concept is very straightforward, it has aroused many controversies. In 1971 Hurlbert arrived to the point to assess diversity as a non-concept: so meaningless that he intimated his colleagues to abandon the term due to the countless conceptual, semantic and technical problems behind it (Hurlbert, 1971). Magurran (1988) successively wrote "[...] diversity is rather like an optical illusion. The more it is looked at, the less clearly defined it appears to be and viewing it from different angles can lead to different perceptions of what is involved.". Nevertheless, measures of diversity are currently the core of ecological studies, with more than 3 million papers on biological diversity (Google scholar results, August 2018). Historically, measures of diversity have been considered to be an index of the system wellbeing and functioning stability (Ptacnik et al., 2008), but also a tool to compare different ecosystems in space and time under different environmental conditions to answer ecological questions. Biological diversity can be expressed at different levels of complexity: genetic diversity, species diversity, functional diversity and ecosystem diversity are all different concepts within the biodiversity bigger picture.

When we consider diversity we have to take into account two main components: the variety of units (e.g., species, functional units), which can be expressed as the number of units (richness), and the relative abundances of each (evenness or equitability or also dominance measures). The real challenge has always been to find a way to combine these parameters in order to obtain a quantification of diversity through the so-called diversity index. You can consider every index as a different point of view of biodiversity, descriptors which are carefully chosen according to the communities and to the ecological question addressed. Every index finds its applicability, exhibiting specific limits

and explicative power (see Hill et al., 2003 for bacterial communities). There is a large number of diversity indexes, each has its limits and applicabilities, but the commonly used are few (Tab. 1.1). Indeed, the most applied indexes are the traditional richness, Shannon-Wiener, the reciprocal Simpson index and the Gini–Simpson index. Magurran (1988) proposed a classification in three groups of diversity measures: richness indexes, indexes based on relative abundances (evenness), and abundance models. Strong positive relationships between species richness, evenness and Shannon-Wiener index have been found in the past (De Benedictis, 1973; May, 1975), together with frequently observed distribution of species abundance following a log-normal distribution (Magurran, 1988a) suggesting a strong coupling between richness, relative abundance and diversity. Following these findings many reviews prefers to use richness as the unique measure of diversity.

In parallel, the functional concept of diversity has attracted exponential interest over the last two decades: it is a component of biodiversity which can be expressed as the value and range of taxonomic and individual traits that influence ecosystem functioning (Tilman, 2001). Some authors (Chapin et al., 2000; Wardle, 2004) suggest that functional diversity could also be a means to decipher the relationship between biodiversity patterns and biogeochemical cycles, providing a new tool to answer ecological questions (Stec et al., 2017). Indeed, the incorporation of phytoplankton functional diversity has proven to be fundamental in the modeling of biogeochemical cycles (Le Quere et al., 2005; Hood et al., 2006). Changes in phytoplankton community structure may have a significant impact on elements cycling at both local and global scales due to the different sensitivities to environmental perturbations of different functional groups (Litchman et al., 2015a). Measures of functional diversity are analogous to taxonomic diversity. Species richness for example finds its functional corresponding in the number of co-occurring functional types (Petchey and Gaston, 2006; Longhi and Beisner, 2010; Behl et al., 2011). Functional types are a group of species aggregated according to their responses

Tab. 1.1: Summary of the most widely used diversity indexes. These same indexes can be applied to measure both taxonomic and functional diversity, the only difference lies in the definition of the units. Whereas for functional diversity the units will be functional types, for taxonomic diversity the units will be a specific taxonomic rank (e.g., species, order, genus).

Diversity index	**Formula**	**Key**
Richness	$R = N$	The richness index corresponds to the total number of units (N) in the dataset. It neglects completely their abundance.
Shannon-Wiener	$H' = -\Sigma \frac{n_i}{N} ln \frac{n_i}{N}$	By contrast, the Shannon-Wiener index weights the number of units present over their proportional abundance ($\frac{n_i}{N}$, with n_i equal to the number of entities belonging to the i^{th} type and N corresponding to the total number of entities in the dataset). According to this index the most equitable community should have an equal abundance of all its units. The more a community has a higher richness and a higher equitability the higher will be its diversity. It includes therefore not only the richness but also the evenness information.
Gini-Simpson	$D = \Sigma \frac{n_i(n_i-1)}{N(N-1)}$	Compared to Shannon, the Gini-Simpson index still include in a unique index both richness and equitability. However, differently from the previous index this latter does not weight the contribution of each units to the index over the logarithm of their relative abundance, giving the same weight to all the units. Consequently, Simpson stands out rarer units, which are almost completely neglected by Shannon.

to the environment as well as their effects on ecosystem functioning (Gitav and Noble, 1997). Conforming to this definition, functional diversity measures should assess the variety of effects organisms have on the particular ecosystem they live in (Cadotte et al., 2011).

In this thesis I will address diatom richness both from a taxonomic and functional points of view, as two straightforward measure of diatom diversity resulting from the two main historical conception of diversity. To have comparable results the whole study is based on data from the same oceanic expedition: *Tara* Oceans (Karsenti et al., 2011, https://oceans.taraexpeditions.org/). Mea-

suring both indexes of diversity gives access both to the distribution from an evolutionary point of view thanks to taxonomic diversity but also to diatoms community stability and dynamics thanks to functional diversity. These two measures provide deeply divergent information and the combination of the two information guarantee thus the complete understanding of the system structuring and evolution.

1.1.2 The plankton paradox and coexistence

The cited large diversity of diatoms, but more generally of the whole phytoplankton compartment, has been source of vast debates among aquatic ecologist because of the discrepancy between the observed high number of phytoplanktonic species at any given time and the low number of essential available resources. Competitive exclusion theory, as articulated by the Gause law, assumes that two species competing over the same resource cannot coexist (Gause, 1934). The popular paradox of plankton (Hutchinson, 1961), stating the observed infraction of competitive exclusion principle, questioned scientist over the last 50 years over the conceptual model of species coexistence in marine systems. A number of hypotheses have been evoked to answer this question (Roy and Chattopadhyay, 2007) including non-equilibrium mechanisms as top-down (Prowe et al., 2012) and bottom-up controls (Huisman et al., 2001), chaos and internal oscillation (Huisman and Weissing, 1999; Dakos et al., 2009) and also mutualistic relationships (Mougi and Kondoh, 2012).

A useful contribution to resolve the plankton paradox are theories linked to resource competition, which would justify diversity on the basis of their resource environment. One of the first author to firstly analyze in a systematic and quantitative way these theories is surely Tilman (1977). According to this theory each species growth depends on the resource concentration it is nourished on, the parameter R. R will be reduced as the population growth

until it reaches the limiting threshold, where the population is no more sustained, and it ceases to grow. At this level, R is denominated critical resource level, R*, where growth rate balances mortality as stated by Eq. 1.1 where r, m and H respectively indicate the maximum growth rate, the half-saturation concentration for nutrient uptake and the mortality.

$$R^* = \frac{mH}{r - m} \tag{1.1}$$

As R* is assumed to be species-specific one nutrient availability is able to explain the equilibrium presence of one species. Increasing the number of nutrients, if different species have different limiting resources they will be able to coexist in this framework. Following, also environmental variability was taken into account: plankton coexists competing for resources in fluctuating environments. Environmental heterogeneity has been hypothesized to work both in time (Robinson and Sandgren, 1983; Sommer, 1989) as well as in space at micro-scale through a patchiness framework (Richerson et al., 1970). If the required resources fluctuates enough in time or space, multiple species can coexist according to niche theory (Richerson et al., 1970; Vandermeer, 1972; Tilman, 1982; Sommer, 1984). Indeed, niche differentiation theoretically lead species to limit their own population, more than they limit other (e.g., by competition or predation/grazing) favouring in this way their coexistence (Chesson, 2000). Later theories suggested that the process of competition itself caused chaotic variations and oscillations in communities structures in terms of relative abundances, promoting the disequilibrium conditions required for species coexistence (Huisman and Weissing, 1999; Huisman and Weissing, 2000). For diatoms in particular, it has been found that a different exploitation of resources between species in the same environment allow their coexistence (Alexander et al., 2015). The two major resources for phytoplankton are nutrients and light (Litchman and Klausmeier, 2008). Resource partitioning could be allowed by species-specific metabolic capabilities and resource availability

responses which results in different transcription modulations. Following this principle, Alexander et al. (2015) found two species typically co-occurring, expressing pathways for the uptake and assimilation of two different pools of dissolved nitrogen. Supporting the same conceptual model, studies focused on nutrient-uptake in diatoms demonstrated highly variable species-specific kinetic parameters (Sarthou et al., 2005; Wilhelm et al., 2006).

Even if bottom-up processes have had large following as biotic control of phytoplanktonic diversity through nutrient availability and seasonality (Dutkiewicz et al., 2009; Barton et al., 2010), niche modeling does not include only abiotic factors and biotic variables as predator (top-down controls) have also been taken into account. Grazer activity may indeed have a fundamental role in defining biodiversity (Hutchinson, 1961; Vallina et al., 2014b; Le Quéré et al., 2015) through different mechanisms. Grazers may randomly predate phytoplankton based on its density, choosing mainly the most abundant taxa and thus limiting the competition winners allowing other taxa to survive and increasing phytoplanktonic diversity. But also different scenarios have been hypothesized, where grazers have preferential preys and predate accordingly, and this latter yields better results in models (Prowe et al., 2012). Hamm et al. (2003) propose the frustule itself as structure specifically evolved to avoid grazing, as a strong force is needed to crush it. This thesis plays in favor of preferential grazing activity as accordingly diatoms may have evolved specific structures as responses to specific predator types. Feeding behaviour can vary between different taxa as well as within the same population according to environmental conditions, indeed it has been foreseen a stronger top-down control in climate change conditions, such as temperature increase (O'Connor et al., 2009; Wilken et al., 2013).

Other non- niche-based explanations of the paradox have been proposed, such as the neutral coexistence theory. According to this latter the high diversity is the result of processes of dispersal limitation, speciation and ecological

drift, allowing coexistence of species also if occupying very similar ecological roles (Hubbell, 2001). This model differs completely from niche theory, starting by the extreme assumption that all species are equivalent between them and having identical impact on one another. This hypothesis hence strides with the major assumption in community ecology theory expecting a positive covariance between the degree of trait similarity of species and their competition (Abrams, 1983). Everything is ruled by stochasticity in these particular models, which is modeled through the randomness of births, deaths and dispersal of species. While niche modeling is based on the peculiarity and unique traits of each species neutrality sinks its roots in species similarities.

At this point a reconciliation of the niche models and of the neutral theory is necessary as both points of view focus on main processes to be taken into account for a "truly unified theory of biodiversity" (Rosindell et al., 2011; Chust et al., 2013). Indeed, we can think of niche and neutral theories as complementary processes controlling community dynamics and for this reason they support each other (Adler et al., 2007). The theory of "lumpy coexistence" is a proposal explanation which can be collocated in this scientific direction. According to this latter, species are self organized in competing assemblages where within each assemblage species are characterized by very similar traits and are treated as nearly neutral (Scheffer and Nes, 2006). Another effort to merge niche and neutrality was proposed by Gravel et al. (2006) with the continuum hypothesis. In their study they modeled the relative importance of niche and neutrality in processes defining community structure and they found that both elements are the ends of a continuum of conditions ranging from competitive to stochastic exclusion.

Within this thesis I will investigate the possibility to assess the environmental niche for diatom putative functional groups. I will thus avoid to define the above-discussed niche of species, hypothesizing an environmental preference of functions rather than taxonomic units. My hypothesis is indeed

that the difficulties encountered in the past to model the environmental preferences of species can be partitioned on the environmental preferences of the corresponding functions. As functions, rather than taxonomic units, are the better units to define the relationship between the organism and the system where it lives in, explaining the environmental preferences of functions could pave the way to more clear and interpretable patterns.

1.1.3 Modeling phytoplankton diversity

Numerical modeling has proved to be a fundamental tool to understand the distribution and diversity of phytoplankton, able to test the different hypothesis behind their regulation parameterizing this information in the model. Looking at the historical development of the conceptual schema beneath models it is possible to ascertain the corresponding development of the ecologists' visions of the planktonic system. We will see in the following how several key elements have been gradually integrated in the models and how phytoplankton itself has been modeled with increasingly complicated structures. Indeed, starting as a unique compartment, phytoplankton can now be modeled by up to 350 different types of phytoplankton, each with its physiology and environmental preferences.

The first phytoplankton models were in the form of NPZ or NPZD configurations, where nutrients, phytoplankton, zooplankton and detritus were the only components taken into account as single elements (Steele, 1974; Wroblewski et al., 1988). To model phytoplankton functional diversity, ecologists firstly introduced the phytoplankton functional types (PFTs) where phytoplanktons are biogeochemically and taxonomically assigned to different functional types according to their role. This approach was developed (Totterdell et al., 1993) to refine the paradigmatic NPZ-type models with a partition of the phytoplankton unit in several functional components. Anderson (2005) reviewed the limits and potentiality of this method: among the future perspectives of

this analysis, he addressed the importance of forward articulate the detail of ecosystem models, deepening the classification of functional units. Discussing and updating Anderson's review (Anderson, 2005), Shimoda and Arhonditsis (2016) compared 124 aquatic biogeochemical models to conclude that the future of PFT modeling is in a gradual increase of complexity. Within this modeling system diatoms are generally characterized by the maximum growth rates among species and high nutrient half saturation (Dutkiewicz et al., 2015). However, life cycle of diatoms is composed of different phases: a growing phase where biomass keep on increasing, a resting stage phase with dormant or quiescent, and finally the phase where they undergo sexual reproduction or they die (Von Dassow and Montresor, 2011). Consequently, each phase is characterized by very different physiology and functional role. For this reason, specific phases should be differentially parameterized within models.

However, increasing complexity and hence the number of functional units taken into account is not an easy task as it amplifies the parametrization effort needed and would lead to an overlapping of trait distribution (Edwards et al., 2015). To answer the demand of a finer definition model a different modeling approach started to consider functional diversity within the diatom group. This was the trait-based approach, where functional groups are defined by key traits describing their function as well as the possible trade-off between them. Different traits can be chosen among morphological, physiological, behavioural and life history information (Violle et al., 2007). Traits are selected to reflect species ecological strategies and indicate how individuals respond to environmental factors, influence other trophic levels and affect ecosystem properties (Kattge et al., 2011). However, the key-traits selection and measure method is the real hindrance of this approach (Petchey and Gaston, 2006). Among the most utilized traits in the literature can be found several examples of traits designed on resource utilization to link ecological processes with biogeochemical processes through species performances (Litchman et al., 2015b). As Litchman et al. (2015; see also Stec et al., 2017) suggested there is

still a need for finer traits measures and 'omic' data has the potentiality to play an important role in the answer (Coles et al., 2017). The integration of omic data within a computational framework would be indeed the new challenge for modelers. This integration could lead to both interpretative and operational models covering different levels of complexity: from individual and population levels to community and ecosystems (D'Alelio et al., Submitted).

The most important point is that models have to be designed upon the specific scientific question or practical aim. There are not and never will be generalistic models, since their complexity would be too unwieldly. In the case of prediction of the response to climate change and other anthropogenic stressors, a new class of modeling approaches to prediction is now starting, a sort of hybrid approach that uses mathematical equations for the parts of the systems that are well understood while using statistical tools such as machine learning for less constrained elements. This is the case of Barton et al. (2016) which exploited machine learning methods to model the niche of phytoplankton species distribution and then combined these models to future predictions of ocean environmental conditions derived from a mathematical model to foresee the future distribution of phytoplankton species.

One step further could thus be to use metagenomics to derive new AI tools specialized on this kind of data and develop new hybrid approaches. In my PhD thesis I'll start on this path exploring the explicative potentiality of AI tools such as Boosted Regression Trees and Neural networks to model the distribution of diatom diversity. The applicability of these methods on omic data opens unprecedented predictive power on a system so complex as the marine one, for which our understanding still does not allow us to resume its dynamics in mathematical expressions.

1.2 Diatoms place within the global ocean

In the following chapter I will present the current knowledge on diatom place within the global ocean. This is to be interpreted as both spatially, in terms of geographic distribution of this taxa, as well as their role within the biogeochemical cycles. In the previous sections I presented diatoms from a biological and ecological point of view and I highlighted the significance of diatom diversity on phytoplankton communities. Hereby this chapter you will find, as consequence of the ecological description depicted in the previous chapter, what is known about the resulting geography of this taxa and how it is connected to global scale cycles of major nutrients, to have an idea of the real weight of diatoms on global dynamics.

1.2.1 Diatoms biogeography

Biogeography, the branch of ecology focused on the species distribution in geographic space, bases its roots in the days of naturalists and early explorers. In the case of plankton, most biogeographies are based on proxies such as chlorophyll (Longhurst et al., 1995) and on seasonality patterns of main forcings (that is, they are based on a bottom up view). In fact, the spatial distribution of phytoplankton, the main primary producers of the whole marine ecosystem, is poorly understood yet, given the limited available data (Maredat; Buitenhuis et al., 2013) and consequently the debate over the factors regulating their distribution is still open. Different theories have emerged over time, feeding threads over the past several decades. The first conceptual schema of phytoplankton distribution was the so-called “Cosmopolitan Paradigm”, summarized by Baas-Becking (Baas-Becking, 1934) through the statement ‘Everything is everywhere, but the environment selects’ (EiE hypothesis). According to this paradigm, micro-organisms may potentially disperse everywhere and reach every geographical region but their local abundance

is determined by local environmental factors. This theory is still source of debate, with some authors (e.g., Finlay, 2002; Fenchel and Finlay, 2004) that support the view of a cosmopolitan distribution of plankton as described in the 'Cosmopolitan paradigm' (Baas-Becking, 1934). However others follow the 'moderate endemicity hypothesis', assessing the possibility of geographical limit in dispersion (e.g., Foissner, 2006; Vanormelingen et al., 2008). Indeed, modeling papers suggest that in some oceanic regions species are actually maladapted, i.e., present in there only because of the transport by the current (Clayton et al., 2013). Recent papers reached the point to support a nearly neutral view of plankton biogeography (Hellweger et al., 2014, for the case of bacteria).

Focusing on the importance of these factors for diatoms, some authors sustain the ubiquity of diatoms, whose community compositions would be predominantly defined by environment's species sorting (Finlay, 2002). Finlay et al. (2002) wrote about the exceptional dispersal abilities of diatoms basing their opinion on the observed ubiquitous distribution of many taxa together with a discussion on the undersampling limits of rare taxa. Cermeño and Falkowski (2009) too, concluded that diatoms' distribution is ruled by the environment while the dispersal represent no limitation to this taxa, supporting the EiE hypothesis. However, several studies suggested endemism, dispersal limitations and geographic limits for diatoms. Conceptually, the authors accepting the existence of microorganisms' biogeography, describe this latter as result of three major drivers: (1) the abiotic conditions, (2) the biotic relationships (i.e., predation, facilitation and competition), and (3) dispersal potential (Follows et al., 2007; Bottin et al., 2016). Among the several examples (Sabbe et al., 2001; Vyverman et al., 2007; Wetzel et al., 2012), Telford et al. (2006) found diatom dispersal to be too slow to overcome the regional metacommunity processes, representing thus an actual limitations shaping the distribution of diatoms. Similarly, Vanormelingen et al. (2008) found diatoms to have biogeography drivers more similar to those of macro-organisms ones, showing

examples of both ubiquitous and endemic diatom's species, supporting the moderate endemicity hypothesis. In this framework, Casteleyn et al. (2010) found a cosmopolitan diatom such as *Pseudo-nitzschia pungens*, to exhibit population differentiation at macrogeographic scales, highlighting the importance of dispersal limitation. Endorsing the same thesis Wetzel et al. (2012) suggested different dispersal capabilities of diatoms, a variability including up to limiting rates. Generally all these studies confirmed the second part of the Baas-Becking statement "the environment selects" (for a review on environmental filtering see Kraft et al., 2015) rejecting however the starting concept expressed by "everything is everywhere", finally articulating as obsolete the EiE hypothesis (Naselli-Flores and Padisák, 2016) toward the 'everything is endemic' philosophy (Williams, 2011). Finally, Malviya et al. (2016) completed the first global study on diatom distribution taking advantage of the unprecedented amount of data supplied by the *Tara* Oceans expeditions. This work finally described diatom distribution at genus level, showing not only that diatoms have specific geographical occurrences but also the relevance of hydrodynamics in their determinations.

Concerning geographies of phytoplankton in diversity terms, similar drivers have been proposed: according to Chust et al. (2013) both the contribution of niche assembly (due to environmental filtering) and drift and dispersal assembly (neutral theory) would describe it. Dispersal plays indeed a fundamental role in shaping and maintaining phytoplankton diversity patterns (Vyverman et al., 2007; Barton et al., 2010; Clayton et al., 2013), allowing coexistence of competitive species (Bracco et al., 2000; Perruche et al., 2010). Lévy et al. (2014) investigated which kind of dispersal arising from different scales of motions have the higher impact on phytoplanktonic diversity among the large-scale circulation, the mesoscale turbulence and the vertical mixing and all three dispersals depicted a significant positive effect on local scale diversity. Higher phytoplankton diversity may be granted also by water dynamics through mixing communities dynamically separated as eddies or water fronts,

which lead to hotspots formation (Lévy et al., 2015). The new challenge of biogeography is to integrate the taxonomic classification of life with a functional one also to better predict global changes consequences for ecosystem functions and services. Functional biogeography has been defined by Violle et al. (2014) as 'the analysis of the patterns, causes, and consequences of the geographic distribution of the diversity of form and function'. The opportunity given by the new global omic databases in this framework is to provide large information from *in situ* data. The same omic dataset can be exploited to investigate taxonomic and functional biogeography. Global patterns that could have been only speculated up to a decade ago are now available for direct observations. In the near future, metagenomics and metatranscriptomic associated to genomic functional marker will provide functional biogeography studies at global scale, towards a better understanding of ecosystem functioning.

After this synthetic overview of the current views and open questions on diatom ecology I can briefly summarize some general traits as follows. Diatoms are known to be very abundant in biomass terms in spring at high latitudes and nutrient-enriched regions such as equatorial and coastal upwelling regions (Tréguer et al., 2018). Periodically they can be important contributors of the phytoplankton compartment also in oligotrophic regions, as mid-ocean subtropical gyres, peculiarly through diatom-diazotroph assemblages (Brzezinski et al., 2011). Diatoms distribution is thus ruled by the ocean dynamics, the availability of nutrients but also their interaction with predators, pathogens and parasites (Tréguer et al., 2018).

1.2.2 Diatoms role within major biogeochemical cycles

> ***Nitrogen***
>
> *Forever cycling*
>
> *from air, to soil, roots, crops, us.*
>
> *Exercise addict.*
>
> **— Mary Soon Lee**
>
> (Poet)

As a consequence of their ecological success, their fast growth rates (Edwards et al., 2015; Flori et al., 2017) and relatively large sizes (Edwards et al., 2015), diatoms from a biomass point of view reach volumes so large to be relevant contributors in biogeochemical cycles. They primarily contribute to the carbon cycle (Smetacek, 1999) but also in the biogeochemical cycling of nutrients such as nitrogen and silicon (Nelson et al., 1995; Armbrust, 2009; Bowler et al., 2010; Tréguer and De La Rocha, 2013). Very roughly, their participation to biogeochemical cycle of nutrients starts from their uptake in the cell and follows in their introduction in the food web through upper trophic levels or alternatively their sedimentation via a sinking process down to the sea floor where the organic matter escapes consumption.

The most studied biogeochemical cycle is the carbon cycle. The process including phytoplankton in this cycle is called the biological carbon pump (Volk and Hoffert, 1985). This mechanism is indeed ruled by the photosynthesis activity of microalgae, which fixes dissolved inorganic carbon into organic carbon and it imports this product into the food web. Grazers rapidly consume the organic carbon produced by diatoms making it accessible to upper layers predators. The fate of this biological carbon, firstly synthesized in the euphotic zone, will fall in one of two options. Most of the biological carbon will be converted back to inorganic CO_2 through food web processes in the upper layers of the oceans but a small fraction will eventually reach the deep ocean,

and once remineralized to CO_2 it will in this way be sequestrated from the atmosphere over geological timescales. Marine diatoms are indeed responsible of around half of marine net primary production (26 TgC yr^{-1}; Conley and Carey, 2015) and they are estimated to fix as much biological carbon yearly as all the terrestrial rainforests combined (Field et al., 1998). Not all diatoms have the same potential as contributor of the carbon cycle: different species have different biomasses and carbon contents which can vary up to nine order of magnitude (Leblanc et al., 2012). The relevant role of diatoms in the biological carbon pump has driven researchers to investigate both the bioremediations application of diatoms algae to fix dissolved CO_2 (Denman, 2008) as well as the response of the global carbon cycle to the increase in atmospheric concentration of the greenhouse gas carbon dioxide. In this context it is fundamental to understand the response of diatoms populations to climate change as alterations or redistributions could have dramatic consequences on Earth's atmosphere (Armbrust, 2009).

Strictly interconnected with the carbon cycle is the silica cycle (Pondaven et al., 2000). The production of biogenic silica in the ocean is mainly attributed to diatoms, while the activity of other siliceous protists (e.g., rhizarians and silicoflagellates) is mostly unknown but it is usually considered to be less significant than the contribution of diatoms (Tréguer and De La Rocha, 2013; Conley and Carey, 2015). Nevertheless, recent works indicate that silicious picoplankton may have a major role in biogenic silica stocks with small but persistent regional contributions, suggesting the need of a revision of diatoms role in this cycle (Baines et al., 2012; Krause et al., 2017). The only soluble form of silica biologically assimilable is orthosilicic acid ($Si(OH)_4$). Diatoms precipitate circa 240 Tmol of silica per year (Tréguer and De La Rocha, 2013). As for carbon (Leblanc et al., 2012), the ability to fix silica differs in quantity according to the species. Recently, attempts to classify diatoms in internal functional groups built on growth rates and degree of silicification were published (Durkin et al., 2012; Assmy et al., 2013). Generally, two classes

have been theorized: a group of small, fast growing, lightly silicified and chain forming diatoms, i.e., the C-sinkers, and a group of large, slow growing, heavily silicified species, i.e., the Si-sinkers. While the C-sinkers are dominant in iron-enriched regions, the Si-sinkers are found predominantly in iron-limited areas.

Another main variable controlling ecosystem properties is nitrogen, whose cycle is often critical for the biogeochemical framework as it is often in short supply relative to the other nutrients required to growth and thus it recurrently is a primary limiting nutrient. Several chemical forms of nitrogen are available in the ocean and their distribution is managed by chemical equilibria. Nitrate (NO_3^-) is the most stable chemical form of nitrogen in the ocean and together with dissolved dinitrogen gas (N_2), they constitute the dominant stock of nitrogen in marine waters. While only Archea and Bacteria can have access to dinitrogen, diatoms, as most of phytoplankton, can assimilate NO_3^- as well as other less abundant sources of nitrogen as ammonium (NH_4^+), nitrite (NO_2^-) and organic compounds such as urea and free amino acids. Nevertheless the most frequently exploited source of nitrogen for diatoms is NO_3^-, the most abundant source at surface and enriched in the euphotic zone through mixing water processes from the stock underneath, but also ammonium, whose recycling is very fast, as it is a mandatory source of nitrogen of all the phytoplankton (Dortch, 1990). Generally NO_3^- is largely more available then NH_4^+ but there are large geographical variations in the supplying of different sources. Due to these many forms, and its strong interconnection and control on other cycles as of phosphorus and carbon ones, the nitrogen biogeochemical cycle is one of the most complex cycles in the ocean, to the point to be addressed as 'deliciously complex' (Zehr and Kudela, 2011). Dinitrogen is 'fixed' (i.e., reduced) by heterotrophic nitrogen fixers: bacteria which are consequently able to release NH_4^+ in decomposition, functioning as nitrogen source for phytoplankton too. Dissolved inorganic nitrogen (nitrate, nitrite and ammonium) is hence taken up by phytoplankton and then integrated

in the food-web through grazers and microbial decomposers' activity. Either entering the food web or directly sinking, leading to the sedimentation of organic detritus to the bottom. The total nitrogen deposition is estimated to be between 46.2 Tg N yr^{-1} (Dentener, 2006) and 67 Tg N yr^{-1} (Duce et al., 2008). Remineralization of detritus accompanied by mixing water processes can bring in surface water new nitrogen sources stocked in the higher depths.

1.3 Diatom nitrogen metabolism

Reading the present introduction up to this point you could understand the relevance of diatom diversity for the global ocean as well as for the biogeochemical cycles in view of climate change (section 1.1). In this optic it likely is the functional concept of diversity the more appropriate approach to understand the relationship of diatom to the abiotic system. Focusing on functional diversity you read about the relevance of the functional trait selection. Among traits, the resource utilization trait emerged as a fundamental one to discriminate phytoplanktonic units. One key step of this thesis is the research of possible molecular marker of the resource utilization trait in diatoms. Being strictly linked to several biogeochemical cycles (section 1.2.2) the choice of different nutrient could be justified but N in particular is recurrently found as primary limiting nutrient and, consequently, it has a leading role in shaping communities (section 1.2.2). Following these principles, in this thesis I investigate the use of gene families from diatom N metabolism as markers. Therefore, in this chapter it is described the current knowledge on diatom N metabolism.

1.3.1 The path from the ocean to the cell

As described in chapter 1.3, nitrogen is supplied in several chemical forms within the marine environment and diatoms have access only to the

sources of dissolved nitrogen. Ammonium and nitrate are the most commonly exploited sources of nitrogen for diatoms, even if ammonium is generally the favored form by phytoplankton. This preference is driven by the reduction of the molecule which translates in a lower energetic cost for its acquisition: indeed NH_4^+ is transported more easily across the cell membrane rather then NO_3^- under N limiting conditions (Glibert et al., 2016). For this reasons, in high concentration of ammonium, the uptake of NO_3^- is inhibited (Conway et al., 1976; Lomas and Glibert, 1999). This inhibition is a consequence of a product of NH_4^+ assimilation and it is not directly activated by the concentration of NH_4^+ in the medium (Syrett and Morris, 1963). This process is highly variable (see the review of Dortch, 1990), however the high level of concentration needed to activate such inhibition are so high (1μM) that it is very unlikely to commonly happen in the open ocean, where concentration are widely lower, but they are conditions rather confined to estuaries, where elevated values ($>$ 5 μM) are now worldwide common (Glibert et al., 2016). Nevertheless, more recently this NH_4^+ preference has been questioned for diatoms. Several characteristics suggest diatoms to be NO_3^- specialists, contrary to cyanobacteria and many chlorophytes and dinoflagellates that are better adapted to exploit NH_4^+ sources (Glibert et al., 2016). Behind this thesis there are several observations: i) diatoms dominate in NO_3^- enriched pelagic environments (e.g., Wilkerson et al., 2000; also seen through enrichment experiments, e.g., Glibert and Berg, 2009); ii) they use NO_3^- even at very high concentration of NH_4^+ (Lomas and Glibert, 1999) with stronger uptake rates and specific growth rates than flagellates at a comparable substrate concentration (Paasche et al., 1984); iii) there is a large phylogenetic distance between diatoms nitrate transporters and other algae nitrate transporters (e.g., Kang and Chang, 2014; Rogato et al., 2015); and iv) the relative higher number of copies of nitrate transporter genes in diatoms compared to other algae (Armbrust et al., 2004). Moreover not only nitrogen preference, availability and inhibition processes play a role in structuring diatoms N utilization but also environmental conditions such as

light and temperature, and species-specificity are important factors to take into account (Song and Ward, 2007; Glibert et al., 2016).

Diatoms are also able to exploit urea and other organic nitrogen forms (amino acids, nucleic acids, and urea), which contribute to around 30% of the global uptake (Allen et al., 2011). Indeed, they own a complete ornithine-urea cycle, absent in plants and green algae but similar to the one present in animals: this metabolic way has been inherited by the heterotrophic host of the secondary endosymbiosis and it has been proven to guarantee to diatoms a fast recovery after prolonged nitrogen starvation (Allen et al., 2011). This cycle may play a role in nutrients transport between the mitochondria, the plastid and the cytoplasm (Bender et al., 2012) but it could also regulate NH_4^+ path in the N metabolism (Armbrust et al., 2004; Allen et al., 2011). Another source of nitrogen can be provided by mutualistic relationship of diatoms together with N_2-fixing cyanobacteria (Hilton et al., 2012). For all the different molecular forms of nitrogen sources diatoms have a specific metabolic pathway of transporters (Tab 1.2) (Armbrust et al., 2004; Hildebrand, 2005).

Tab. 1.2: Diatom nitrogen transporters.

Nitrogen source	Transporter
Urea	Urea transporters (URT)
NH_4^+	Ammonium transporters (AMT1)
NO_3^-	High/Low affinity nitrate transporters (NRT2/NPF)

Nitrate transporters are classified in high and low affinity transporter systems, according to the affinity and capacity in transporting nitrate. Both low affinity nitrate transporter (NPF or NRT1) and high affinity nitrate transporter (NRT2s) perform an active symport transport proton-coupled (H^+) prompted by pH gradients across membranes (Navarro et al., 1996), although some evidence suggest the use of Na^+ rather than H^+ in this symport mechanism for marine diatoms (Rees et al., 1980; Boyd and Gradmann, 1999). NRT2's substrate is NO_3^- but it may also transport NO_2^- while for NPF a large variety of substrate were reported in plants: NO_3^-, di/tri-peptides, amino acids, dicar-

boxylates, glucosinolates, auxin (IAA) and abscisic acid (ABA) (Frommer et al., 1994; Liu, 1999; Jeong, 2004; Waterworth and Bray, 2006; Krouk et al., 2010; Kanno et al., 2012; Nour-Eldin et al., 2012). Specifically for diatoms, NPF transporters ability is still to be analyzed but phylogenetic similarities suggest their ability to transport dipeptide (Rogato et al., 2015). High affinity ammonium transporter (AMT1) is a channel-like protein acting as NH_4^+ uniporters or NH_3/H^+ cotransporters. The two processes of uptake and assimilation are often uncoupled in phytoplankton, evidence of this phenomena are the large intracellular nitrate pools found also in diatoms (Lomas and Glibert, 2000). The degree of this uncoupling is species-specific and related to growth conditions (Colios, 1982; Dortch et al., 1991) but recent knock-out experiments in diatoms proved the complete lack of communication between uptake and assimilation of nitrate in diatoms too (Mccarthy et al., 2017).

Once inside the cell, in the cytosol nitrate is reduced to nitrite by the assimilating enzyme nitrate reductase (Galván and Fernández, 2001; Allen et al., 2005; Bowler et al., 2010). The produced NO_2^- is then transported in the chloroplast: both in the chloroplast (Galván et al., 2002) and in the cytosol (Armbrust et al., 2004; Allen et al., 2006) nitrite is further reduced to ammonium by nitrite reductase (NiR), respectively in the forms of Fd-NiR and NAD(P)H-NiR. Ammonium (directly obtained from uptake or produced by NO_2^- reduction) is at this point assimilated by glutamate synthase (GOGAT) /glutamine synthetase (GS) to produce glutamate, essential for amino acids biosynthesis (Takabayashi et al., 2005). The assimilation of ammonium in diatoms can be located in the plastids (Fd-GOGAT / GSII) as well as in the mitochondria (NAD(P)H-GOGAT / GSIII) (Bowler et al., 2010; Allen et al., 2011) where this process may be activated by ammonium produced by cytosolic catabolic reactions (Parker and Armbrust, 2005; Hockin et al., 2012). The relationship between nitrogen and carbon metabolism is very tight: the connection between the two paths occurs at the level of amino acid biosynthesis, where the tricarboxylic acid cycle (TCA) supply the reducing equivalents and

carbon skeletons needed (Hockin et al., 2012). A summary of the described metabolic pathway can be observed in Figure 1.1.

1.4 Meta-omic as new tool to study ecological questions

1.4.1 Meta-omic: a new powerful tool of investigation yet to be properly exploited

Sequencing of DNA from environmental samples without any culturing step (Handelsman et al., 1998) is defined as metagenomics and it allows us to investigate how micro-organisms are structured, how they interact among them, and adapt to their environments (Tyson et al., 2004; Allen and Banfield, 2005; Gill et al., 2006). Due to the recent advances in environmental DNA sequencing and the drop of its cost, a number of metagenomic project arose in the last decade to decipher microbiomes of different habitats (Raes et al., 2011). As the production of this massive amount of datasets keeps on going, the challenges shift now to the downstream computational analysis (Falony et al., 2015). While metagenomics is defined as the sequencing of the environmental DNA, that is, the whole-community DNA, metatranscriptomics (cDNA) is its complementary information. The first identifies the genes and pathways present in a community whereas the second depicts their expression. These approaches allow the assessment of the genomic and transcriptomic composition and diversity within and across different communities, along with targeted rRNA gene sequencing (16S in bacteria and 18S in eukaryotes).

Even if this revolutionary approach promised to unravel complex ecosystem functions of micro-organisms biome (Simon and Daniel, 2011) we still miss the methods to fully understand the data we are producing (Stec et al.,

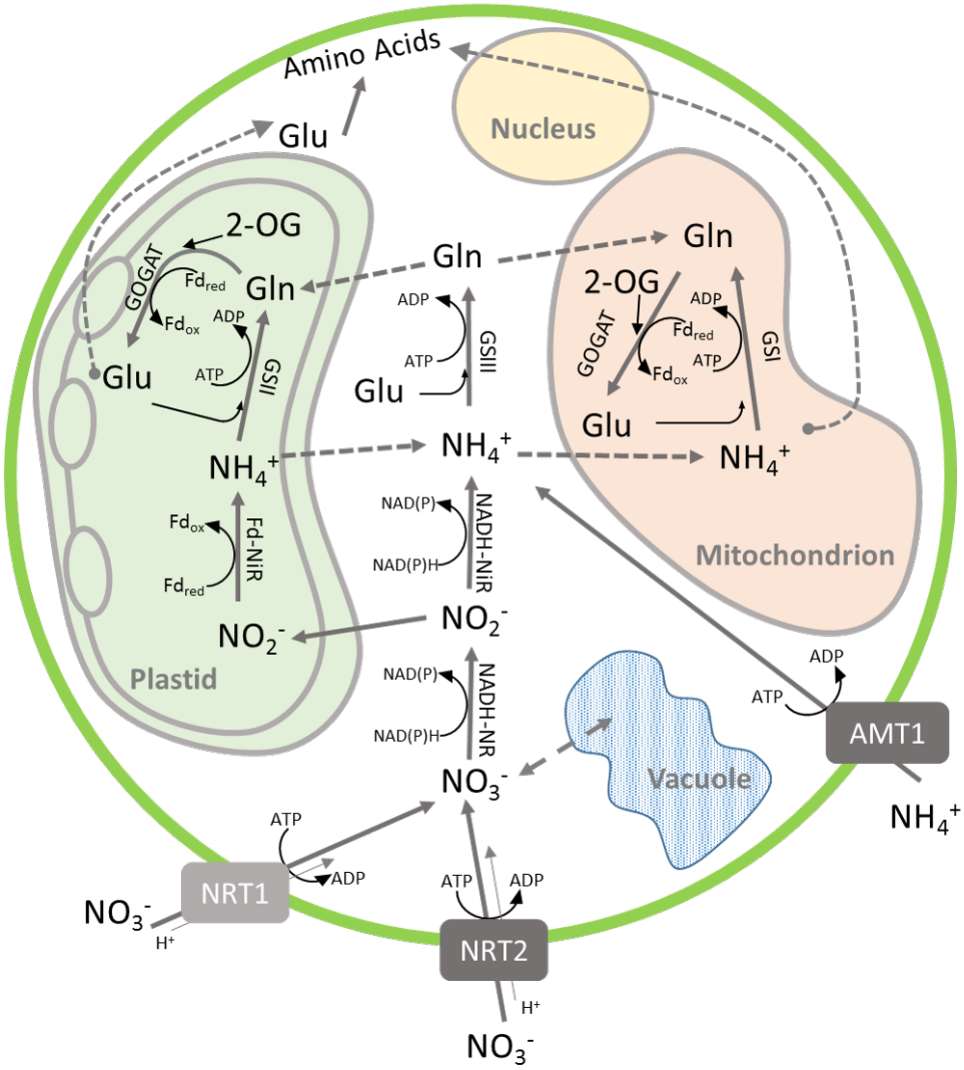

Fig. 1.1: Conceptual schema of the major processes within diatom nitrogen metabolism related to the uptake and assimilation of nitrate and ammonium. Typically the reduction of nitrate to nitrite occurs in the cytosol, while the following reduction of nitrite to ammonium occurs in the chloroplast. The assimilation of ammonium have different possible locations: the cytosol, the chloroplast or the mitochondrion (inspired from Glibert et al., 2016).

2017). Metagenomic sequencing reaching saturation is supposed to allow the assemblage of the complete genome of all the organism in a community at the theoretical level (Hess et al., 2011; Narasingarao et al., 2012). However, reaching the sequencing saturation is not always obvious and the majority of the paper published on even high impact journal does not reach it (e.g., Tringe et al., 2005; Gill et al., 2006; Sun et al., 2015). Nevertheless, even if in most cases it is impossible to obtain whole-genome assembly from these data, good reconstruction of the most abundant micro-organisms from metagenomics has proved to be reachable even without meeting all the previous cited methodological objectives (Segata et al., 2013) through reference based (e.g., AMOS), or de novo methods (e.g., MetaVelvet-SL, SPAdes and IDBA-UD). But limitations are not confined to the sequencing operations.

Bioinformatic analyses allow us to infer general patterns ruling microbial ecosystems but we still have not designed analyses able to completely describe microbes ecology and evolution within their environment (Hiraoka et al., 2016). Most of the analyses developed requires references, depending on straightforward searches on the genomic dataset based on sequence similarity against the known references. Taxonomical assignation for example, which is one of the main first steps in describing microbial community, is based on sequence similarity searches against reference genomes (e.g., RefSeq) or 16S rRNA sequences databases (e.g., Greengenes, SILVA, RDP and Ez-Taxon). As straightforward as it may seem taxonomic assignation is still very challenging: because of the limits of reference databases, as well as the difficulty to discriminate two closely related species, or the ambiguities derived by the presence of intraspecific variability, conserved regions or the results of horizontal gene transfer products (Pignatelli and Moya, 2011; Mende et al., 2012). The reference limits are probably the most dramatic of this approach hinders, to the point that using different datasets may produce different taxonomic annotations (Pignatelli et al., 2008). The enrichment of databases is thus a fundamental missing piece now not only to increase the number of reference

but also to smooth the bias toward model organisms. For this reason, several projects are striving to obtain new genomic sequences to enrich existing databases (e.g., Yamazaki et al., 2009; Grigoriev et al., 2014). An alternative to these methods are reference-free approaches which cluster reads to group marker genes, as 16S rRNA, to define unique representative sequences used as operational taxonomic units (OTUs). Another possible path for metagenomic analysis is to focus on metabolic pathways rather than taxonomic assignations, annotating metabolic genes through sequence similarity searches against pathway databases such as KEGG, MetaCyc and SEED. Investigating the metabolic processes encoded in the genome is a direct information of the microbe response to the environmental conditions they live in. In Figure 1.2 you can find a schematic summary of the different approaches for omic analysis.

1.4.2 How can it be put at the service of ecology?

Among the first communities analyses through culture-independent genomics approaches there are marine microbes (Giovannoni et al., 1990). These methods are fundamental for microbial ecology as the majority of environmental microbes have been found to be uncultivable (Rappé and Giovannoni, 2003; Narihiro and Kamagata, 2013) but focusing on the marine context we also have to take into account the difficulties in collecting samples of specific areas. These communities count a large heterogeneity of phyla including viruses, archaea, bacteria and eukaryotic species. Organisms are distributed differently on size terms and for this reason, studies usually filter the samples in different size classes to focus on specific groups. Different size classes are indeed dominated by different organisms and according to the objective the research targets specific size-classes. But what are we able to infer through meta-omics datasets? Are these the proper tools to answer ecological questions? Even if the level of complexity is so high that we still have to develop the right informatic tools to ascertain all the information (Stec et al., 2017; chapter 4.2), the research community agrees on the incredible

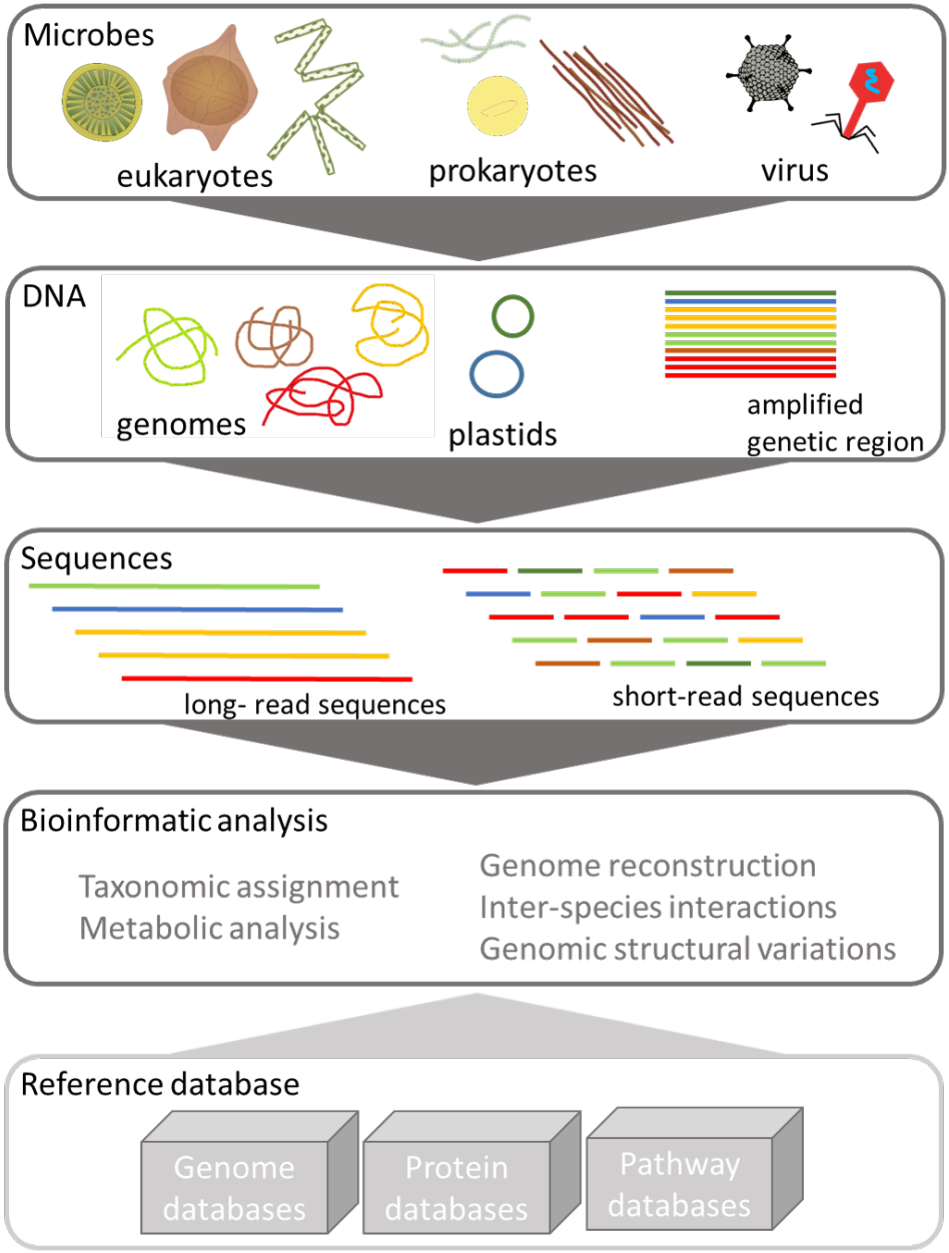

Fig. 1.2: Schematic representation of metagenomic and bioinformatic analysis applied in microbial ecology (inspired from Hiraoka et al., 2016).

potential of this data for microbial ecology defining it the 'most unrestricted and comprehensive approach' (Gilbert et al., 2011).

The first large study on microbial ecology through omics approach has been the Global Ocean Survey (GOS): a metagenomic transect survey from the Northwest Atlantic through the Eastern Tropical Pacific (Rusch et al., 2007). This expedition has been a pioneer project in aquatic ecology but not only, paving the way for large scale ecological surveys in the ocean (Bork et al., 2015) but also in the soil (Vogel et al., 2009; Gilbert et al., 2014) and the human-related biomes (Turnbaugh et al., 2007; Nelson et al., 2011).

Ecologically, this kind of data allow us to describe the community both in taxonomic terms and in metabolic terms, having access both to the diversity and activity of the community, as well as the presence of specific micro-organisms of interest or metabolic pathway or individual genes. The coupling of metagenomic data with contextual data derived from *in situ* measurements, modeling or previous information on the sampling site, gives us the opportunity to infer the relationship between the community and environmental drivers. But the analysis can follow through several directions. Among these, recently studies are proposing finer-scale study of genomic structural variations (e.g., Smillie et al., 2011). Another possibility is to focus on biotic interactions, both at level of inter-species interactions, like mutualism and parasitism, through co-occurrence and network analyzes (e.g., Chaffron et al., 2010; Beman et al., 2011; Biswas et al., 2015; Krohn-Molt et al., 2017).

1.4.3 *Tara* Oceans

Following the footsteps of great ocean explorers *Tara* Oceans consortium planned an interoceanic sampling campaign to sample the ocean through the latest molecular technologies challenges: the metaomics. The success of this €10 million public/private expedition spans from scientific dissemination, with

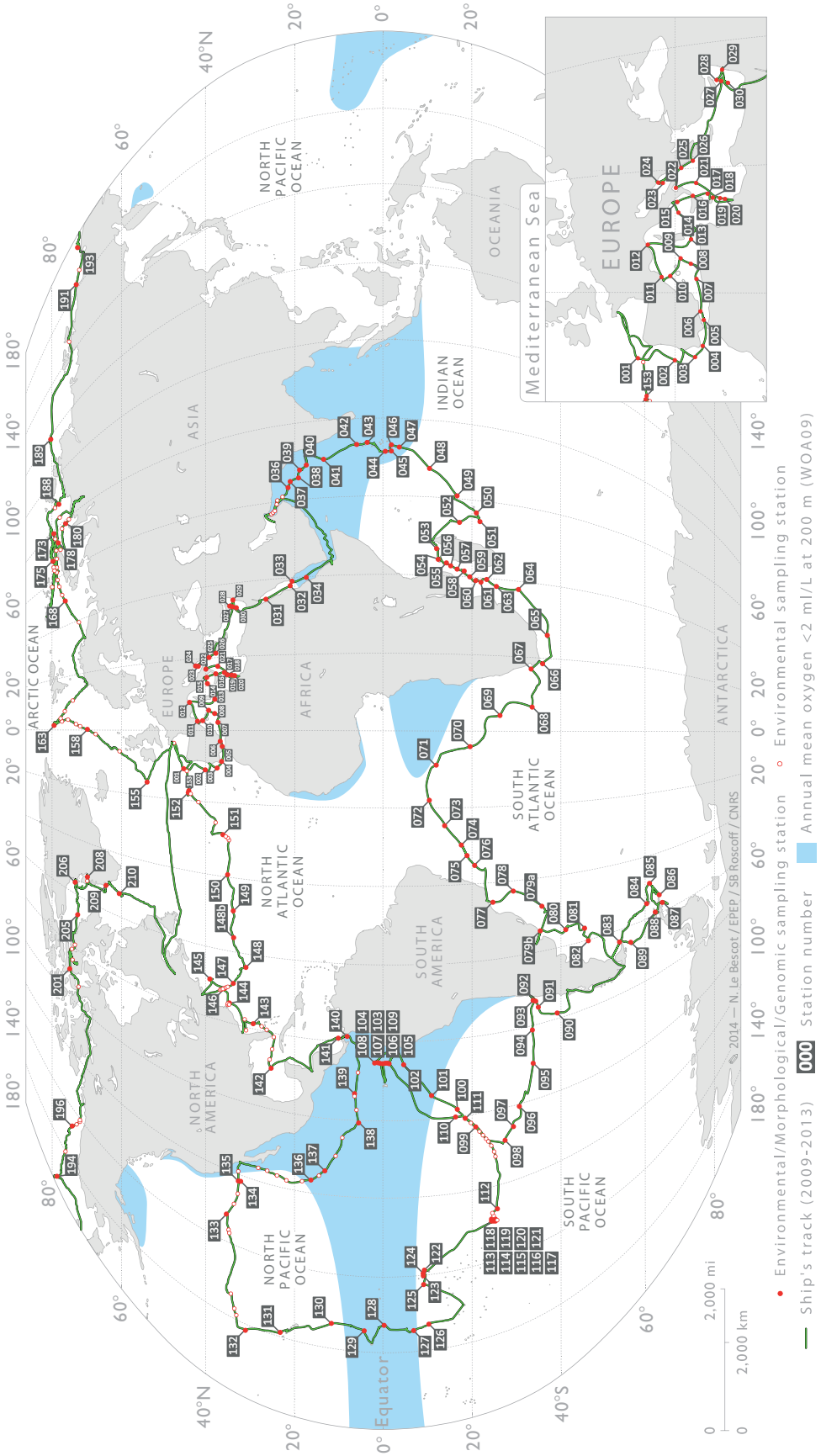

Fig. 1.3: Map of *Tara* Oceans sampling from Pesant et al., 2015.

countless inclusion of kids from schools all over the world and the presence of several artists and journalists on board over several navigation stretches, to true scientific relevance and innovation thanks to the production of the most complete oceanic metagenomics and metatranscriptomic dataset ever built. This project has seen the cooperation and daily effort of a team of over 500 international scientists from 40 nationalities.

The schooner *Tara*, just 36 meters-length, crossed the oceans from 2009 through 2013 sampling 210 sites (Fig. 1.3) and depths up to 2'000 m covering seven oceanographic provinces: the North Atlantic Ocean (NAO), Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO) and South Pacific Ocean (SPO) (Karsenti et al., 2011). Each station has been carefully placed according to oceanographic structures, weather and season to capture a single picture of specific conditions such as mesoscale eddies, upwellings, anaerobic zones, main currents and acidic waters, chosen for their expected biodiversity or climatic relevance. The placement of the stations and sampling depth was carried daily based on *in situ* data, satellite and literature information.

Mainly four size classes have been analyzed from 0.2 to 2'000 μm to study the entire planktonic life from viruses to prokaryote and eukaryotes. For every station up to three depths have been sampled: a subsurface layer (5 meter depth, SRF), a mesoscale depth and the deep chlorophyll maximum (DCM) depth. In addition to genomics purposes, many samples were taken for other more traditional analysis like microscopy identifications, imaging, and flow cytometry. Together with the biotic data a large variety of environmental information were taken, with around 30 directly measured parameters to describe the sampling conditions (Pesant et al., 2015). Totally, *Tara* Oceans sampled ~35'000 biological samples and ~13'000 contextual measures.

All size-fraction samples underwent DNA extraction and consequent sequencing library preparation for metagenomics analyses. Metagenomic libraries were prepared manually or in a semi-automatic manner (Alberti et al., 2017).

The metatranscriptome dataset has been built over 68 different sampling sites selected across all the oceanic provinces except the Arctic in both the euphotic zone and the DCM. The four size fractions, independently sampled were included in the analysis. High-coverage polyA-based RNA-Seq was implemented in a total of 441 plankton communities, reaching 16.5 terabases of raw data. Consequently cDNA reads were assembled per sample and then clusterized together at 95% sequence identity to compile a catalogue of 116.8 million sequences of at least 150 bases. These sequences have been called unigenes and they will be hereafter termed like this in the present manuscript. Interestingly, only 48% of the unigenes could be taxonomically annotated delineating the large percentage of unknown organisms within the communities. All the details on the metatranscriptome dataset can be found in Carradec et al. (2018).

The *Tara* Oceans dataset has been already exploited to study diatoms. Previous studies described the taxonomic abundance and diversity at global scale (Malviya et al., 2016) and found out that nanodiatoms are much more relevant on the community than was previously appreciated (Leblanc et al., 2018). Moreover using the same dataset diatoms have been found to be antisocial (Lima-Mendez et al., 2015; Hendry et al., 2018) and to live in symbiosis with other organisms (Vincent et al., 2018).

1.5 Aim of the study

The general aim of my thesis is to explore global diatom diversity at different scales from the taxonomic to the functional. For this purpose I exploited different meta-omics data, proposing, all along the thesis, the integration of multivariate analysis and machine learning approaches for the investigation of this kind of data. This study was based on the global-scale *Tara* Oceans datasets, including the metagenomic, the metatranscriptomic as well as the metabarcoding datasets.

In Chapter 2 I have investigated diatom richness from a taxonomic point of view. Given that measures of this information largely diverged according to metabarcording or morphology-based measures I developed a pipeline able to combine the two kinds of information to obtain an optimal filtering of metabarcoding-derived data. Once this technical challenge was overcome I obtained a new measure of taxonomic diversity and used it to address fundamental ecological questions such as the definition of its distribution, the environmental processes underlying it, and an analysis of the conditions allowing hotspot formation. By exploiting AI approaches I assessed the nature and relevance of the environmental cues triggering diversity, in order to better understand the conceptual model behind the maintenance of species coexistence.

After a thorough analysis of diatom taxonomic diversity, the subsequent three chapters assess diatom functional diversity by exploiting the metatranscriptome and metagenome datasets within the same *Tara* Oceans framework. Moving from a taxonomic to a functional measure of diversity I go deeper into the ecological stability of diatom communities and on their impact on the ecosystem. Within this context I propose a first attempt to measure diversity in functional terms purely on marker gene phylogeny selected as descriptors of a single functional trait.

In Chapter 3 I chose to focus on the resource utilization trait and I selected two N uptake gene families as molecular descriptors. Their phylogeny allowed the determination of putative functional units, assessed in this pipeline as the evolutionary clades of N transporter genes. Within this same chapter I go deeper into the taxonomic composition of single clades to understand the functional distribution across diatom phylogeny.
In Chapter 4, by exploiting presence-absence data of N transporter clades I could derive information about putative functional richness. I studied the biogeography of this diversity index and consequently I looked for the environmental drivers of this diversity.
To conclude the study on diatom putative functional richness, in Chapter 5 I investigated the modulation of the functional units. To do so I took advantage of the abundances of these units as derived from the corresponding mRNA levels in the metatranscriptome dataset. Herein, I thoroughly searched for possible relationships with the ambient environmental conditions and the utilization of each single functional unit.

Finally in Chapter 6 a very preliminary but promising exercise has been included. Through this thesis I investigated diatom functional diversity on a global scale by exploiting the unprecedented omics data derived from the *Tara* Oceans project. Before having access to these global scale datasets the only means to have an idea of global patterns was through modeling approaches. These methods developed through the last decades have improved step by step, reflecting more and more processes of the ocean system. From very simplified models, the number of phytoplankton units included in the models are increasing, up to the recent model of Dutkiewicz et al. (unpublished data) which includes 350 phytoplankton types. The strong limit of these models is the difficulty to compare them to the real data. Through this chapter I search for correspondences between the 350 types derived by the model and the *Tara* Oceans metabarcode OTUs. This assessment is one step toward the integration of modeling and omics data, giving more introspection both on the

identity of the phytoplanktonic types included in the models as well as on the parameterizations used to obtain units corresponding to specific OTUs, giving insights into physiologies of species that never entered a wet lab.

2

Taxonomic richness of diatoms resolved by different measures

2.1 Summary and main achievements

- In this chapter I provided a filtering-based reconciliation between metabarcoding and morphology-based estimations of diatom richness exploiting the *Tara* Oceans dataset.

- I was able to reach a significant correlation between the two richness estimations with a coefficient up to 0.62 applying a specific filtering pipeline over the metabarcode dataset;

- The filtered pipeline herein designed focuses on the exclusion of the rarest OTUs. It finds the optimal reconciliation with the microscopy data through the exclusion of the rarest OTUs, responsible of the cumulative relative abundance of 0.35% of the total abundance.

- The filtering application produced a great change in metabarcode-based richness patterns at global scale affecting the most polar stations, the more enriched in diatoms in abundance terms;

- Exploiting the filtered richness dataset, I confirmed the unimodal relationship between biomass and richness already proposed by the literature;

- I studied the differences between the filtered and unfiltered metabarcoding datasets, analyzing the percentage of filtered OTUs and the lost on phylogenetic diversity across the stations. The filtered OTUs may be artifacts and/or strains, population related to more abundant OTUs present in the sample. Further efforts should be addressed in this direction to correctly assess the nature of these OTUs.

- I built a model able to efficiently predict diatom richness exploiting hydrodynamic and nutrient data. This model is a great achievement of the thesis due to the high accuracy of its estimations, its only weakness lies on the detection of diversity hotspots.

- I propose a combination of mechanicistic (hydrodynamics) and bottom-up processes as main players in hotspots formation, having different relative importance at the local scale, and I investigate the processes behind the *Tara* Oceans diatom richness hotspots through a neural network approach.

2.2 Introduction

The study of phytoplankton diversity is of particular interest for understanding the factors ruling ecosystems stability and functioning (Ptacnik et al., 2008). While biodiversity patterns have been widely investigated in the terrestrial ecosystem we still lack enough data to reach a comparable understanding of planktonic communities. Phytoplanktonic biodiversity indexes are based on taxonomic identification and counts of the observed organisms. Among the different ecological indexes developed to describe biodiversity, richness (expressed as the number of species present in the system) is surely the most straightforward (Magurran, 1988a). Taxonomic richness is basically just the count of species present in a site, even if the concept is truly basic, it still is

an index very sensitive to the sampling effort and dramatically affected by rare entities (Cermeño et al., 2014). The problem subsists mainly because of the incredibly large quota of rare species within planktonic communities (Sogin et al., 2006; Caron and Countway, 2009; Ser-Giacomi et al., 2018). The composition of marine microbe communities is indeed typically characterized by a few dominant species largely abundant and an incredibly high number of rare species. This rare community has been named 'the rare biosphere' (Sogin et al., 2006). Logares et al. (2014) for example found a striking fixed percentage of rare OTUs over their samples of around 70% of the whole detected OTUs. This great amount of rare species has a cumulative low weight on the total abundance but it makes the difference in terms of diversity indexes such as richness, but also phylogenetic diversity, of which they are responsible of the most contribution. This same rare component can be seen also as a large diversity reservoir, built of ecologically redundant species, able to rapidly interact with the environment and eventually easily change the community structure (Caron and Countway, 2009). To account for these rare species two contrasting approaches have been developed: either you estimate the true richness thanks to species accumulation curves and rarefaction analyses, or you neglect the rare species by filtering all the elements under a certain threshold of abundance. An example for the second case is the already cited Vallina et al. (2014) which applied a 1% threshold. In this kind of systems the latter threshold means to capture only the very abundant organisms, a very small percent of the total number of species. Clearly, the results of different approaches lead to highly different estimations of planktonic richness.

These conceptual limits are strictly related and consequent to the methodology applied to measure diversity and the corresponding resolution. Indeed, traditional methods for diatom identification are based on the use of light microscopy and morphology-based approaches exercised by specialist taxonomists. To face both the limits of microscopy identification in terms of resolution and human errors but also the time and money expenses which

implies, molecular approaches such as next generation sequencing (NGS) combined with DNA barcoding have recently become a very attractive alternative solution. Metabarcoding is a means to obtain a standardized and cost-effective characterization of microbial communities (Comtet et al., 2015). This method exploits short gene sequences as markers in order to identify species. These markers are selected because they are ubiquitous in the tree of life and are characterized by the presence of highly variable regions that are diagnostic of different taxa. These latter are used to distinguish different species flanked by very conserved regions used to design the primers to amplify the variable regions: typically 16S SSU rDNA gene (V4 region) is used as marker for bacterial communities whereas 18S SSU ribosomal DNA (V4 or V9 region) are exploited to study eukaryotic ones. Of note, an ongoing discussion focuses on the optimal choice of DNA barcoding markers for diatoms. Among the proposed markers, beyond the more common 18S SSU rDNA (Zimmermann et al., 2011; Luddington et al., 2012), several DNA barcoding markers have been discussed: the 28S LSU rDNA (D2/D3 region) (Hamsher et al., 2011), *cox1* (Evans et al., 2007; Evans et al., 2008) and several sequences within *rbcL* (Hamsher et al., 2011; MacGillivary and Kaczmarska, 2011; Kermarrec et al., 2013). Metabarcoding has proved to be of great potential for diatom identification (Kermarrec et al., 2014; Visco et al., 2015; Zimmermann et al., 2015; Kelly et al., 2018). The application of molecular approaches could also pave the way towards the discovery of new diatoms species (Mann, 2010). The comparison of microscopy and metabarcoding methods for this taxa has proved highly reproducible for NGS together with an overall greater analytical precision (Zimmermann et al., 2015; Malviya et al., 2016; Kelly et al., 2018). Microscopy identification lacks resolution and is unable to reliably document rare species. Moreover, morphology-based measures of phytoplankton richness are usually underestimated by sampling effort constraints (Cermeño et al., 2014) to the point that the use of rarefaction curves or richness estimates are strongly encouraged for this kind of data (Gotelli and Colwell, 2001). Nevertheless, metabarcoding has its own limits: it can consider absent species

actually present (false negatives) but it can display false positive too. False positives may be produced by contamination, PCR or sequencing errors (Ficetola et al., 2015; Ficetola et al., 2016). Other constraints are linked to taxonomic identification due to the lack of a complete and thoroughly curated reference catalogue, which leads to a lack of resolution in OTUs identification, if not a lack of identification at all. Finally, V9 or V4 are not fully resolving taxonomically neither for diatoms nor for other plankton groups (e.g., Dunthorn et al., 2014). As both have their advantages and limits Groendahl et al. (2017) suggested the integration of the two methods to assess the community composition of eukaryotic microorganisms. The reconciliation of the two methods is definitely one of the current challenges on marine ecology (Muller-Karger et al., 2018). The high resolution of metabarcoding is appealing for the fundamental role played by rare species. Addressing thus the problem of rare species detectability by different methods we forcily interact with the portion of rare biosphere we can observe and the possible reservoir we measure. A further source of information are metagenomic and metatranscriptomic data, which may be potentially more informative than those exploited to date. Notwithstanding, like the other kinds of data these also are likely to suffer from limitations and a proper methodology. Indeed, a proper measure of richness from such kinds of data has not yet been developed. We thus need to reach a consensus on the correct way to estimate and measure the richness index, a step strongly needed in particular for the planktonic world, where rare units, difficult to measure, may be key descriptors of the system.

But what do we expect to obtain, by measuring phytoplanktonic richness? First, richness is a measure of the *mean* properties of the community, because if plankton are able to recurrently bloom it means that they may be always there or be recurrently transported there, at least in the open ocean. Thus, this is in contrast to the use of the Shannon index of diversity, which emphasizes the distribution of the abundances across the species and thus is an almost instantaneous property, given that the planktonic system has time scales of

response of only few weeks. Indeed, the richness index reflects not only the throphic status (e.g., oligotrophy etc.) but also the time variability of the environmental conditions since it measures also the pool of species that are dormant or not active, waiting for their optimal season (see below).
Secondly, since the time scale of populations (months to years) are similar to that of the ocean physics (mixing, dispersal transport by main currents) we can expect local communities to be regulated by ecological (local adaptation) and physical processes (transport of populations) (Clayton et al., 2013). To have an idea of the importance of currents, one can consider that the typical velocity of ocean currents is 10-20 cm/s (8-17 km/day), meaning that, during a typical bloom duration (two weeks) a population will typically move at about 150-300 km/day. In the case of the Gulf Stream, water velocity is over 100 cm/s and thus more than 85 km/day (i.e., 1,200 km in two weeks). During the journey the population will undergo significant mixing with neighboring populations, living in different environmental conditions. It is thus evident that, at a specific site, the populations are the result of a combination of ecological and physical factors, where the importance of the latter depends on the degree of the local strength of the oceanic mixing and transport. A typical example is the strong mixing at confluences (D'Ovidio et al., 2010). In turn, this implies that it is difficult to assume that a single factor or process shapes the local richness everywhere. It is thus more reasonable to assume that different factors have a different importance depending upon their local relative importance. For this reason it is also important to use analytical tools for the analyses that are conceived for emphasizing this locality of the ruling processes.

Up to now, many efforts have been made in several directions to characterize phytoplankton diversity relationships with respect to i) biomass (or productivity), ii) geography and iii) the biotic and abiotic processes influencing it.

Regarding the association between phytoplanktonic diversity and biomass, studies have found a unimodal relationship, with higher diversity corresponding to an intermediate phytoplankton biomass or productivity for both taxonomic (Irigoien et al., 2004; Passy and Legendre, 2006; Spatharis et al., 2008; Vallina et al., 2014a) and functional (Török et al., 2016) richness. Santos et al. (2015) found taxonomic richness to be better than other diversity indexes such as the functional ones for predicting phytoplankton productivity. This peculiar unimodal association is taxa-specific, being for example more emphasized in diatoms compared to dinoflagellates (Spatharis et al., 2008). Indeed, even if this relationship is visible also in *Prochlorococcus*, *Synechococcus* and flagellates, whereas for the first two the unimodal peak of richness is shifted to lower-intermediate primary production, for flagellates and diatoms richness indexes peaked towards higher-intermediate primary production (Vallina et al., 2014a). Different hypotheses have been proposed to explain these particular shapes: at low biomass levels phytoplankton richness may be controlled by nutrients (Spatharis et al., 2008) and/or selective grazing (Vallina et al., 2014b), while at high biomass concentrations the same index may be reduced by bloom conditions (Spatharis et al., 2008) and/or its consequent competitive exclusion processes (Vallina et al., 2014b). Nevertheless, interpreting the origins of this curve is not an easy task as it is bidirectional and each variable is both the cause and the consequence of the relationship (Worm and Duffy, 2003).

Geographically, there seems to be a typical gradient of several indexes of diversity with latitude. This pattern, firstly observed for terrestrial organisms and lately partially confirmed for marine organisms, exhibits a unimodal or bimodal curve, always depicting a decrease of diversity at higher latitudes (Hillebrand, 2004; Barton et al., 2010; Chaudhary et al., 2016).

Finally, for phytoplankton in particular, biogeographies of diversity have been described by both the contribution of niche assembly (environmental filtering) and dispersal assembly (neutral theory) (Vergnon et al., 2009; Chust

et al., 2013), as reviewed in chapter 1.1.2). Noteworthy, niche assembly is not to be considered only in abiotic terms, and several biotic processes have to be taken into account (Gray, 2001). Indeed, both the bottom-up view, driven by nutrient availability and seasonality controls (Dutkiewicz et al., 2009; Barton et al., 2010), and the top-down one, fueled by selective and unselective grazing (Hutchinson, 1961; Vallina et al., 2014b; Le Quéré et al., 2015), have been discussed as main players in phytoplankton diversity regulation.

The lack of a univocal definition of diversity, the difficulty in sampling the ocean in different times of the year, the limitations of modeling in reconstructing it (see Barton et al., 2010, where the modeled richness is about one or two orders of magnitude smaller than the actual observed one) and, above all, the technical difficulty in reaching a true estimate of all the species present at a given site at a given time makes the debate on plankton diversity still fragmented and somewhat immature.

In this context, the aim of this chapter is to provide a global scale analysis of diatoms richness dynamics and its relationship with biotic and abiotic variables, investigating the processes behind hotspots formation. To describe such distributions I propose a reconciliation of diatom richness as derived from two methods: the metabarcoding based on the V9 region in the small ribosomal unit (18S) of rDNA and the light microscopic identification, both based on the *Tara* Oceans expedition data. *Tara* Oceans data has already proved the potential of metabarcode analysis for several taxa (De Vargas et al., 2015; Le Bescot et al., 2015) as well as specifically for diatoms (Malviya et al., 2016). Herein, I study the distribution of diatom richness focusing on the local processes shaping it and particularly determining which conditions allow the formation of taxonomic hotspots.

2.3 Material and Methods

2.3.1 Data

Tara Oceans barcoding

Diatom metabarcoding data for *Tara* Oceans samples for the size fraction 20-180 μm has been exploited for the present chapter (Supplementary material File 1). A total number of 194 samples encompassing 135 stations sampled at subsurface (5 m) and 59 of them at Deep Chlorophyll Maximum depth (DCM) as well were used. Total nucleic acids (DNA+RNA) were extracted from all the samples, and the hyper-variable V9 region of the nuclear 18S rDNA was amplified through PCR (De Vargas et al., 2015). A quality filtering based on reads quality checks and a minimum number of occurrences of three copies in at least two different samples was implemented to reduce PCR- and sequencing- errors (De Vargas et al., 2015). Please look at Vargas et al. (De Vargas et al., 2015) for details on the sequencing protocols of the V9 sequences.

Within these stations 237,565 V9 diatom-assigned unique amplicons were found, of which 136,163 were found present in size fraction 20-180 μm (Supplementary material File 2). Out of these latter, 66,372 amplicons were detected only once across the *Tara* Oceans stations (Fig. 2.1) Over this same dataset single OTUs have abundances from a minimum of 4.1e-5%, up to 85% of the whole diatoms OTUs abundances on the sample.

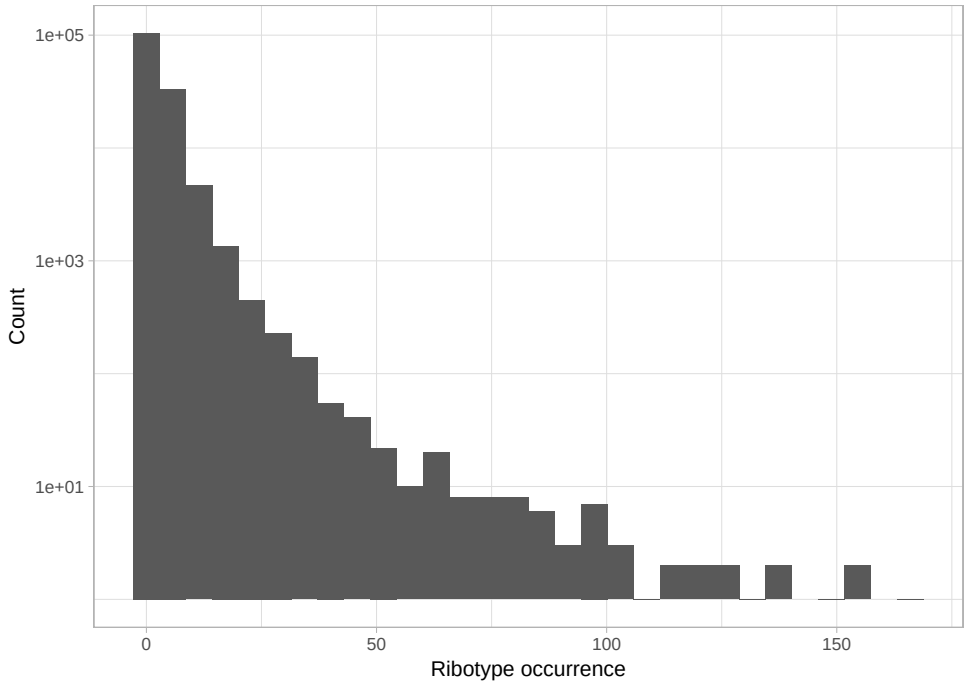

Fig. 2.1: Histogram of number of detections of single amplicons within the *Tara* Oceans samples at the 20-180 μm size fraction.

Morphology-based counting

Morphological-based counting was implemented for the 20-180 μm size fraction samples at surface and DCM depth from the Cape Agulhas region (stations 52, 64, 65, 66, 67, and 68), the South Atlantic transect (stations 70, 72, 76, and 78), the Southern Ocean stations (stations 82, 84 and 85), and South Pacific Ocean stations (stations 122, 123, 124, and 125) (already published in Malviya et al., 2016). But also for other sampling stations, still unpublished data: stations in the Mediterranean Sea (station 7), in the Indian Ocean (stations 36, 38), off the Argentinian waters (stations 80 and 81), in the Pacific Ocean (stations 93, 100, 102, 109, 126, 127, 128 and 135) and in the Gulf Stream region (stations 144, 145 and 146). Three ml of each sample was placed in an Utermohl chamber with a drop of calcofluor dye (1:100,000), which stains cellulose. Cells falling in 2 or 4 transects of the chamber were identified up to the species level if possible and enumerated using a light inverted microscopy (Carl Zeiss Axiophot200) at 400x magnification. This

data is attached as Supplementary File 2.
The identification and enumeration of phytoplankton was performed by Dr. Eleonora Scalco from Stazione Zoologica Anton Dohrn.

Phylogenetic analysis

Phylogenetic analysis were performed on the Swarm metabarcoding d1 clustered at theshold equal to 100 and 99.65 (see following methods) using the approximately-maximum-likelihood method (Yang, 1994) and implemented in the FastTree2 software (Price et al., 2010). The phylogenetic tree was built by Dr. Luigi Caputi from Stazione Zoologica Anton Dohrn.

2.3.2 Swarm clustering

Unique sequences were clustered into operational taxonomic units (OTUs) applying the Swarm approach (Mahé et al., 2014). This method uses 1 base pair difference (insertion-deletion-substitution) steps between barcodes to aggregate them. I computed Swarm aggregation at different clustering levels (d) from 1 to 5, using the standard values for all the others parameters through the SWARM software (Mahé et al., 2014).

The Swarm clustering was performed under the supervision of Dr. Roberta Piredda from Stazione Zoologica Anton Dohrn.

2.3.3 Filtering process

For every station I sorted the OTUs abundances in a decreasing order and then I discarded the least abundant ones on the base of a percentage threshold on the cumulative abundance. This is not equivalent to exclude

OTUs with abundances below a certain relative abundance threshold (e.g., Vallina et al., 2014b) as I always remove the same percentage of abundance from the samples. I prefer to refer to the excluded cumulative percentage since this has different impacts on different samples according to their rank-abundance curve. The same exercise was performed considering thresholds that exclude the cumulative abundance of the rarest species from 1 to 100% of the total abundance, with a step equal to 0.05. This was repeated using unique sequences and Swarm OTUs at 5 different clustering levels, that is with a total of 6 different barcoding information (from d1 to d5, the more stringent clustering threshold). The morphology-based richness was computed as the number of different taxa identified in the samples. Then, pairwise Pearson correlations of the richness computed over the six different metabarcode information filtered at different thresholds were performed against the morphology-based richness. A 'fdr' p-value adjustment of the p-value was implemented. The best correlation between morphology-based richness and any barcode-based richness at different thresholds was selected as the significant correlation (adjusted p-value<0.05) with the higher ρ. The whole filtering procedure is resumed in Fig. 2.2, and the diatom richness estimation at different thresholds computed by the 6 barcoding information are found as Supplementary File 4. From here onwards I refer to the filtered richness as the metabarcoding richness filtered at the optimal threshold from the selected clustering threshold.

The filtering threshold corresponding to the best correlation was implemented over the whole *Tara* Oceans dataset and the obtained richness mapped over a global map.

2.3.4 Biomass and richness

The richness obtained by the unfiltered and filtered dataset were compared against a proxy of diatoms biomass: the relative abundance of diatoms over the sample expressed as the percentage of barcode units assigned to

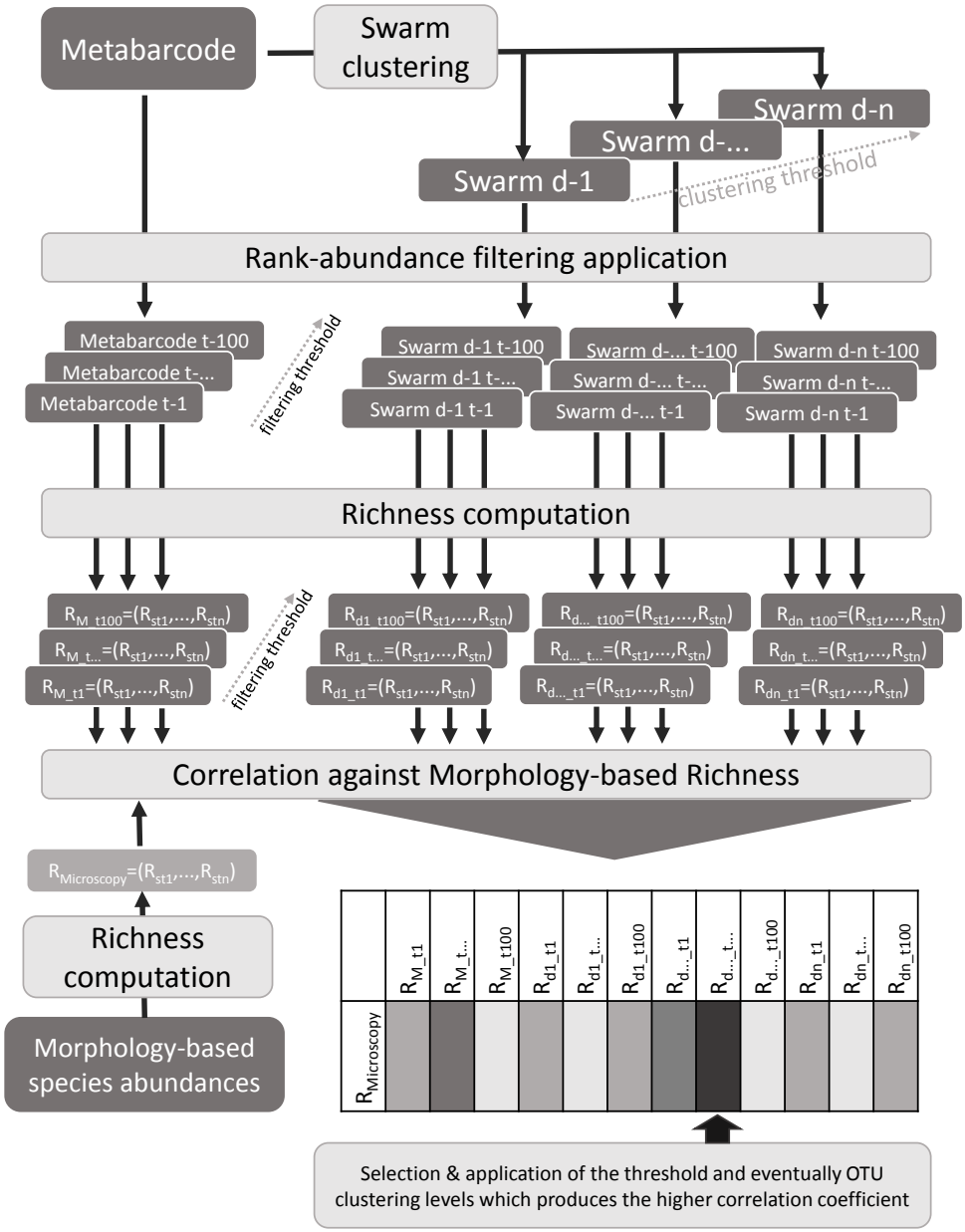

Fig. 2.2: Conceptual scheme of the filtering process. Firstly the metabarcode clustering is developed to obtain from the unique sequences the Swarm OTUs, applying 5 different clustering thresholds (from d= 1 to d= 5). The unique sequences and the differentially clustered Swarm OTUs are then filtered applying different cumulative thresholds on the relative abundance, (from t= 0% to t= 100%). Then, the richness index is computed on all the resulting datasets, to obtain the diatom richness for every sample (from st_1 to st_n). In parallel, richness is estimated based on the morphology-based counts. Finally, a correlation is implemented between the morphology based richness and the ones obtained from unique sequences and from differentially clustered Swarm OTUs, at different filtering thresholds. The optimal metabarcoding dataset, the optimal filtering threshold and eventually the optimal clustering threshold are selected to be the ones providing the best correlation. The best correlation is the statistically significant correlation exhibiting the highest Pearson ρ.

diatoms. Over the resulting scatter plots a fitted line was plotted through the 'loess' method for the two different filtering thresholds datasets.

2.3.5 Filtered OTUs

The number of OTUs filtered out was measured as the percentage of OTUs kept from the optimal threshold filtering. Following, this information data was mapped on the global ocean and compared to the relative abundance of diatoms over the samples. Phylogenetic diversity (PD) per sample has been computed over the Swarm d1 OTUs both on the filtered (optimal threshold = 99.65) and unfiltered data. The number of OTUs present in the 20-180 μm size fraction are of 8,884 and 3,531 respectively for the unfiltered and filtered data. PD was computed through mothur (Schloss et al., 2009) using the normalized count tables (i.e., the sum of all the OTUs taken into account in the unfiltered dataset is equal to 1) and the corresponding phylogenetic trees.

The PD was computed with the supervision of Dr. Roberta Piredda from the Stazione Zoologica Anton Dohrn of Naples.

2.3.6 Boosted Regression Tree

To investigate which environmental variables have a role in shaping diatom biodiversity, a boosted regression tree (BRT) model (Elith et al., 2008) was run using as diatom richness information (the predicted variable) the Swarm d1 metabarcoding filtered at t = 99.65 measured in surface. This approach combines the complex fitting capability of a regression tree with boosting, an adaptive method for combining many simple models to give improved predictive performance. BRT provides several information other than the model itself: that is i) how much single predictor variables are used by the model to predict the predicted variable (i.e, contribution), an

information which gives insights over the influence of each variable over the predicted one, and ii) the response of the predicted variable to single predictor variables that can be interpreted as the univariate niches of the variable of interest (see below). The advantages of BRT are that it can fit complex non-linear relationships, it is adapt at avoiding overfitting (De'ath, 2007), and it can accommodate any type of variable as predictor (continuous, categorical, also missing and non-independent data). Models were applied through the *dismo* and *gbm* R package (Ridgeway, 2006; Hijmans et al., 2017). To better understand the processes structuring in the model only surface samples were included in the BRT model. This choice is justified by the preference of homogeneity of conditions among analyzed samples. Limiting the machine learning analysis to surface samples allows to specifically investigate the processes typical of this depth, while much more complex would have been integrating also the processes of DCM. Not only these particular structures are subject to completely different controllers of diversity but they are present only in stratified water columns and also very different between them being localized (and accordingly sampled) at completely different depths. The methodological approach would allow the inclusion of both sampling depths but to ease the interpretation I preferred to proceed with only surface samples.

Environmental parameters

Nine predictor variables were selected as descriptors of processes behind diatom richness dynamics covering variables related to mixing, nutrient availability variables, temperature and chlorophyll α abundance. I selected as estimates for the local convergence and mixing the finite-size Lyapunov exponent, which is a measure of the relative divergence in time for a given distance of transported particles and it is evaluated via an altimetry-based Lagrangian computation, and, as estimate of the intensity of the local frontal structures,

the local horizontal gradient of sea surface temperature (SST), measured through satellite observations (D'Ovidio et al., 2010). Both parameters were computed by Dr. Francesco D'Ovidio (CNRS, France) in the context of a collaboration. For the nutrient availability the choice fell on iron, silica and nitrogen availability, the latter in the forms of ammonium, nitrate and nitrite. Measures of these data at 5 meters depths were extracted by the Darwin-ECCO2 model for iron and by World Ocean Atlas for the nitrogen sources. Only the silica availability was computed *in situ* (Picheral et al., 2014b). While for iron *in situ* measures were not performed during the *Tara* Oceans expedition, several nitrogen measurements were taken *in situ*. The preference of modeled data was driven by the preference to avoid missing values in the predictor descriptors. It is thus to note for the following that iron and nitrogen contents refer to the local mean state of the ocean and are not suggestive of the instantaneous state during the sampling time.

Parameterization

The BRT model has been run with a Laplace distribution (Kozubowski and Podgórski, 2012). BRT requires the selection of two main parameters: the 'learning rate' and 'tree complexity'. The learning rate defines the contribution of each tree to the whole model while the tree complexity controls whether interactions are fitted: a value of one fits an additive model, a value of two fits a model with two-way interactions and so on. I selected slow learning rates (0.005) able to build models estimating reliable responses and tree complexity equal to 5, to include complex interactions.

Models were run using a 50% bag fraction and a 10-fold cross-validation. A k-fold cross-validation procedure was used to train (90%) and test (10%) each model and select the optimal number of trees. The significance of the model has been estimated through a coefficient of determination, expressed as

the Pearson correlation coefficient between the observed and predicted diatom richness for the same *Tara* Oceans stations.

To understand the factors that determine the richness distribution I then estimated the relative importance, or contribution, of the predictor variables to the model. Contributions reflects the importance of each covariate in the BRT model. This value is expressed as number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees. The contributions are then scaled to sum to 100 where a higher number indicates a higher influence of the predictor on the model structuring.

Sensitivity exercise

I chose the BRT approach since it can consider contributions to the prediction that derive from non-linear relationships between factors and can be easily used to evaluate the local importance of each factor. The latter is important since, as explained above, here I do not assume that these factors have the same importance everywhere. A sensitivity exercise has been implemented to understand the importance of each predictor in the model for each sampling station and to test if my hypothesis of a local dependency was correct. The model has been run nine further times, excluding each time one of the environmental variable used as predictors. The prediction thus obtained was compared to the prediction of the original model and to the observed richness abundances. This comparison aims to understand for every station if the model without the variable improved or get worse compared to the original model prediction. For each station i the prediction quality improvement as influenced by the predictor X is:

$$Q_i = \frac{|R_{pred-TOT} - R_{observed}| - |R_{pred-excludedX} - R_{observed}|}{R_{observed}} \quad (2.1)$$

Consequently a negative value of Q_i means that the parameter X is locally important to the model in station i.

2.3.7 Self-Organizing Map

I applied a self-organizing map (SOM), a type of neural network (see below), to explore the linear and non-linear relationships between diatom richness in surface samples as derived from the metabarcoding filtering and two sets of environmental parameters: hydrodynamical and nutrient ones.

Environmental parameters

As for the BRT modeling (2.3.6) the hydrodynamic parameters set includes the finite-size Lyapunov exponent and the gradient of surface water temperature. Nutrient availability set of variables included iron, and nitrogen availability, the latter in the forms of ammonium, nitrate and nitrite, all derived by global ocean models (DARWIN-ECCO2 for iron and WOA for nitrates concentrations). Herein, the preference of modeled data was driven by the inability of the SOM to deal with the several missing value that characterize *in situ* data.

SOM

This neural network approach analyses the variance structure of the data set and clusterize the items, in our case the sampling stations, identifying the

relationships between descriptor variables without the rigid assumptions of linearity or normality. Each cluster is represented by a unit called node within the SOM: a (usually two-dimensional) grid displaying all the nodes ordered according to the similarity of the model beneath each node (Kohonen, 2001). Indeed, more similar models are associated to adjacent nodes, whereas less similar models are located farther away from each other in the grid space. This kind of organization allows an insight of the relationships of data through a topographic representation. SOM characterizes the models of each node through the weight of each explicative variables, here the environmental parameter, in the model itself.

I applied supervised mapping, through the xyf-SOM network to understand how the diatom richness is influenced by different environmental parameters. The supervised mapping choice allows a more powerful modeling choice in complex data situation (Melssen et al., 2006). The XYF network build two matrices: one using the environmental parameters and the second using a categorization of the richness value in four richness levels. It then concatenates these two matrices training on the similarity of nodes in both. This ability to incorporate the richness information in the environmental-based SOM is the real attribute of xyf-SOM (Melssen et al., 2006). The result of the xyf-SOM approach is two concatenated maps: one on the richness classes and one other with the environmental variables, but the topology of nodes is shared between the two.

Analysis were performed through the R package *Kohonen* (Wehrens and Buydens, 2007). I weighed the X and Y spaces to have equal weights one over the other, I set the number of iterations for the xyf-SOM training process (rlen) to be 100 and the learning rate α started from 0.05 and linearly decreased to 0.01 over rlen updates as set as default (Wehrens and Buydens, 2007).

2.4 Results and Discussion

2.4.1 Filtering process

The *Tara* Oceans barcode counts a number of 237,565 unique sequences annotated to the Bacillariophyta group. The total number of OTUs produced by the Swarm clustering at different thresholds is reported in Tab. 2.1. The richness estimated through the unique sequence numbers and the different clustered Swarm metabarcodes found in the size fraction 20-180 μm was thus compared to the richness obtained by morphology based identification within the same size fraction through a series of Pearson correlation run with several levels of filtering threshold.

Tab. 2.1: Number of Swarm OTUs produced by different clustering thresholds.

Clustering threshold	Number of OTUs
d1	14,480
d2	7,854
d3	5,416
d4	4,042
d5	3,103

The correlation between the same Swarm d1 dataset and the morphology based one with the abundance-filtering passing is only ρ=0.43 at threshold=100 (i.e., for the unfiltered case). The best significant correlation was observed using as omic information the Swarm metabarcode at clustering threshold d=1 filtered at an abundance threshold equals to 99.65, where the correlation peaks at ρ=0.62 (Fig. 2.3). This is suggestive of the power of Swarm clustering in reaching operational taxonomic units closer to the ones used on morphology-based measures (i.e., species). However, this also means that discarding just the rarest OTUs corresponding to the cumulative abundance of only 0.35% of the total abundance of each sample lead to the best reconciliation between the unique sequences and metabarcode dataset together with the (much more expensive and time consuming) microscopy-

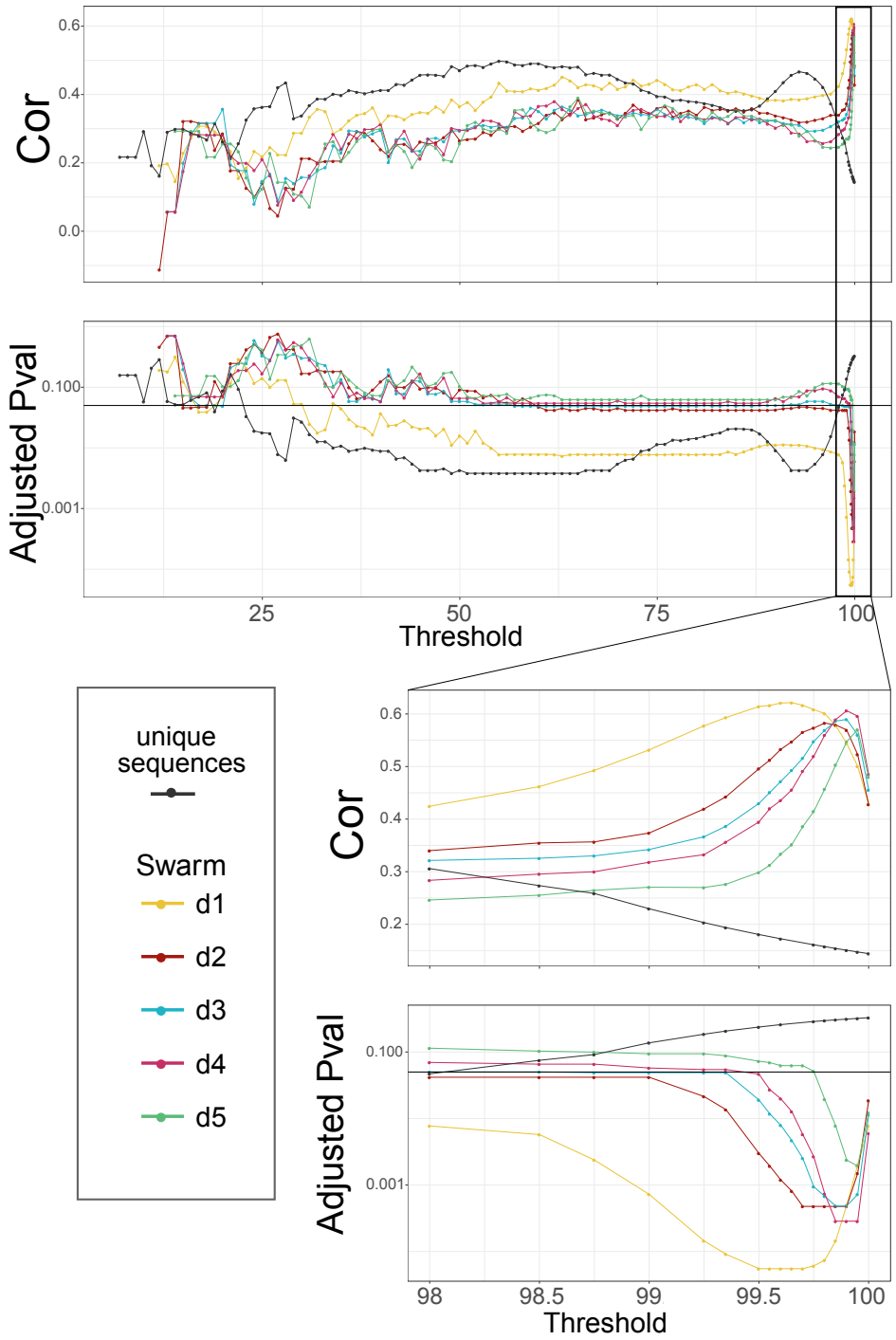

Fig. 2.3: Correlation between diatom richness computed from unique sequences or Swarm metabarcode datasets and from microscopy observations (fraction 20-180 μm). Each panel has an upper panel where the Pearson ρ correlation is shown, and a lower panel with the corresponding adjusted p-values. Correlations are calculated by progressively filtering out rare OTUs. For example a thresholds of 95% means that only the most abundant OTUs making the 95% of the total abundances are kept. The lower panel is just a zoom of the 98%-100% threshold range.

based one. As we will see in the following, this tiny correction gives completely different pictures of the global diversity patterns.

2.4.2 The reconciled diatom richness

Surprisingly, the use of the full metabarcode information or the (very slightly) filtered dataset produced two very different global richness patterns. The unfiltered Swarm metabarcoding showed a strong severance between high latitudes stations and the others, depicting strong peaks of richness in the first ones (Fig. 2.4A). A different landscape is depicted by the application of the same filtering threshold to all the *Tara* Oceans Swarm samples available (Fig. 2.4B). I obtained through this method a biogeography of diatom richness with maxima at mid-latitudes and in the Tropical Pacific. Peaks in diatom richness are observed in the Antarctic, as well as off the well-known upwelling region caused by the Benguela Current (off southern Africa) and the Humboldt Current (off Peru and Chile) and also in correspondence of a bloom in the Marquesas Islands (stations 123 and 124). It is important to note that the range of richness observed after the filtering (Fig. 2.4B) is just ¼ of the one observed on the unfiltered dataset (Fig. 2.4A), however it is still 4 times that observed by the morphology-based approach (Fig. 2.4C). Thanks to the integration of the microscopy based information (Fig. 2.4C), the latitudinal contrast of richness has thus been dramatically altered, with similar patterns but very different gradients (Fig. 2.4B). The latter indeed resembles (and thus apparently validates) the patterns of phytoplankton (not only diatoms) richness recently modeled (Fig. 2.4D; Vallina et al., 2014a). Note that I have filtered out species with relative abundances of 0.35%, that is much less than the threshold used by Vallina, equal to 1%. Using the same value as the cumulative threshold (and not as univocal threshold as applied by Vallina et al., 2014a) I obtain a correlation of 0.53 (Fig. 2.3) and a pattern that is quite different (data not shown). Of course the maximum richness observed is lowered (from a richness equal to 288 at t=99.65 to a richness equal to 174

at t=99) but again, the impact is not equally distributed in terms of richness. In this case a drop in richness is observed particularly in the Indian Ocean and the South Pacific Ocean indeed, depicting quite a different scenario. This latter highlights the strong impact that different filtering thresholds can have in the computation of a diversity index such as richness, strongly sensitive to the large quota of rare units directly touched by any filtering method.

From the applied filtering pipeline the most affected stations were the polar ones. Indeed in these stations I observed richness comprised between 750 and 1,000 OTUs (Fig 2.4A), while once filtered no stations reached a richness higher than 300 OTUs. Looking at the number of OTUs retained from the filtering process in percentage over the unfiltered richness highlighted this latitudinal difference (Fig. 2.5B). While high latitude stations retained only 0-30% of the original Swarm OTUs observed as present, in the middle latitudes values the percentage of OTUs kept from the filtering move to 70-100%. It occurs that the stations more sensible to the filtering, the ones hence with a longer 'tail' of rare species in the rank-abundance plots, were the most rich in diatoms in terms of relative abundance over the sample. An example is depicted in Fig. 2.6 where a polar station and a tropical station rank abundance plots were compared. It is clear the difference between the two curves: the tropical station has a higher slope and a shorter tail, whereas the polar station, even if it has a very high number of OTUs present, exhibits a very long tail. The shape of rank-abundance curves is a diversity indicator of the population structures, and while Ser-Giacomi et al. (2018) did not find any geographic pattern studying the whole plankton compartment I detected a strong difference between polar and non-polar stations while focusing only on diatoms. Comparing the impact of the filtering process, expressed as percentage of OTUs retained from the filtering, to the relative abundance of diatoms in the samples unveiled the linkage between the two (Fig. 2.5A). The most affected stations, the polar ones, are also the stations characterized by the highest relative concentration of diatoms in the samples.

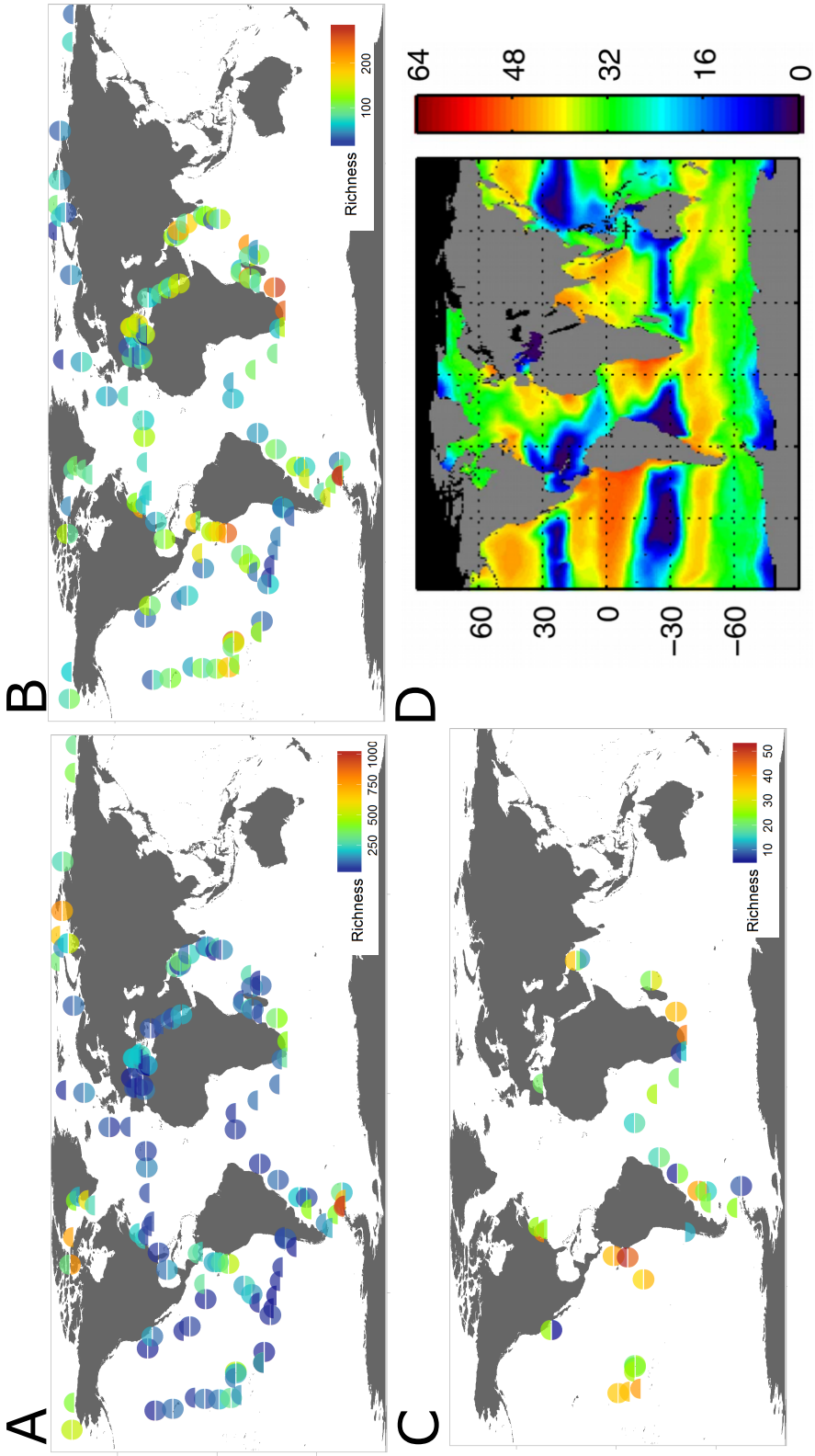

Fig. 2.4: Maps of diatom richness derived from different datasets. In panel A) and B) the information is derived from the Swarm metabarcoding in size-class 20-180 μm respectively unfiltered and filtered at threshold= 99.65 on the cumulative abundances of Swarm OTUs. In panel C) diatom richness is measured from morphologic observations (i.e., microscopy counts) from the net samples of size-class 20-180 μm. In panel D) phytoplankton richness as modeled by Vallina et al. (Vallina et al., 2014a) as number of species contributing > 1% to total biomass.

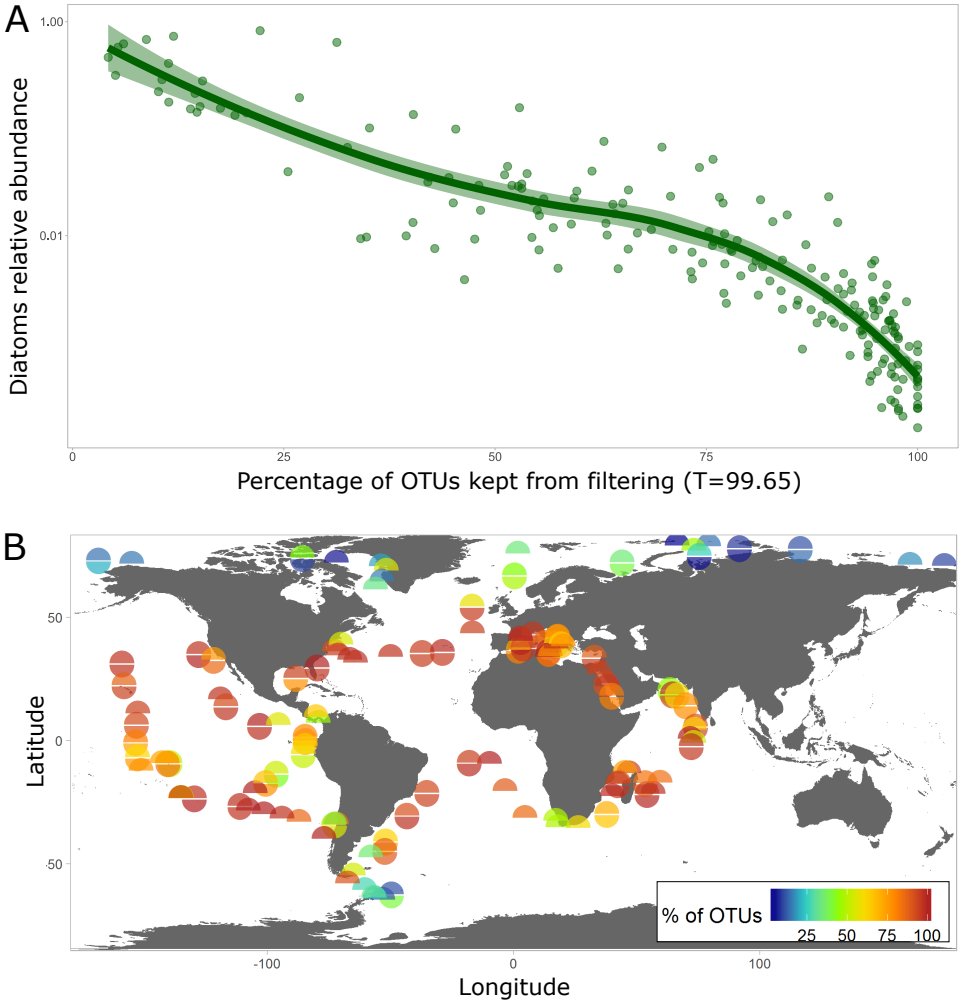

Fig. 2.5: In panel A) the relative abundance of diatom Swarm OTUs over the Swarm metabarcode samples is compared to the percentage of Swarm OTUs kept over the filtering with a threshold of 99.65. The higher the relative abundance of diatoms in the samples the higher the number of Swarm OTUs filtered out from the filtering method. In panel B) a map of the percentage of Swarm OTUs retained over the filtering with a threshold of 99.65 over the Swarm metabarcode (20-180 μm). Stations from high latitudes are the ones with a longer 'tail' in the rank-abundance plots, the ones with higher relative abundances of diatoms in the samples and consequently the more affected ones from the filtering approach we propose.

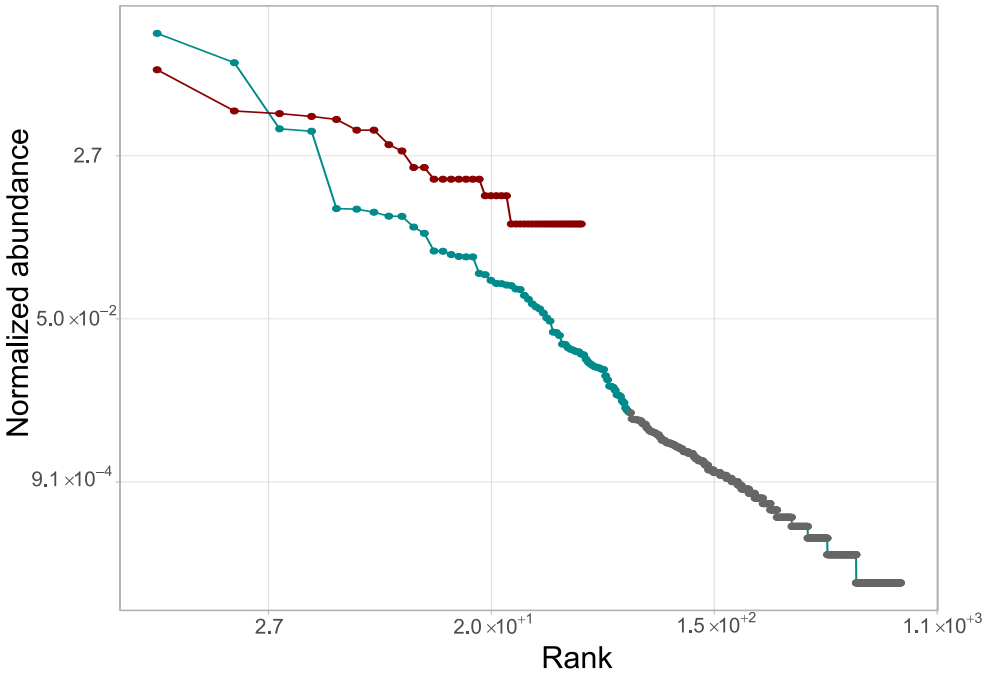

Fig. 2.6: Rank abundance plots for surface samples of a polar *Tara* station in cyan (#188) and a tropical *Tara* station in red (#143). The abundance is normalized, and the sum of all the OTUs abundances in a station is equal to 100. The rarest OTUs, excluded by the filtering process at threshold equal to 99.65, are depicted in gray.

Consequently, I investigated the relationship between OTU richness and the relative abundance of diatoms in the samples. While there is a linear relationship between the relative abundance of diatoms and the observed diatom richness in the unfiltered Swarm metabarcode (Fig. 2.7A - orange), we observe a bell-shaped relationship between the two parameters once the datasets have been filtered (Fig. 2.7A - cyan). This bell-shaped relationship is comparable to the one obtained by Vallina et al. (2014) (Fig. 2.7B) using the diatom relative abundance as a proxy of their biomass. Interestingly, as the cited work from Vallina et al. (2014) depicts the specific productivity-richness curve for diatoms with a peak of richness at slightly higher values of biomass compared to other taxa, my results show similar trends (Fig. 2.7B). This assumption is based on the fact that diatoms tend to dominate in richer environments. Using chlorophyll α as biomass proxy gives a similar result, indeed. This unimodal relationship between phytoplankton diversity and its biomass has been observed by several studies (Irigoien et al., 2004; Passy and Legendre, 2006; Spatharis et al., 2008; Vallina et al., 2014a) and the fact

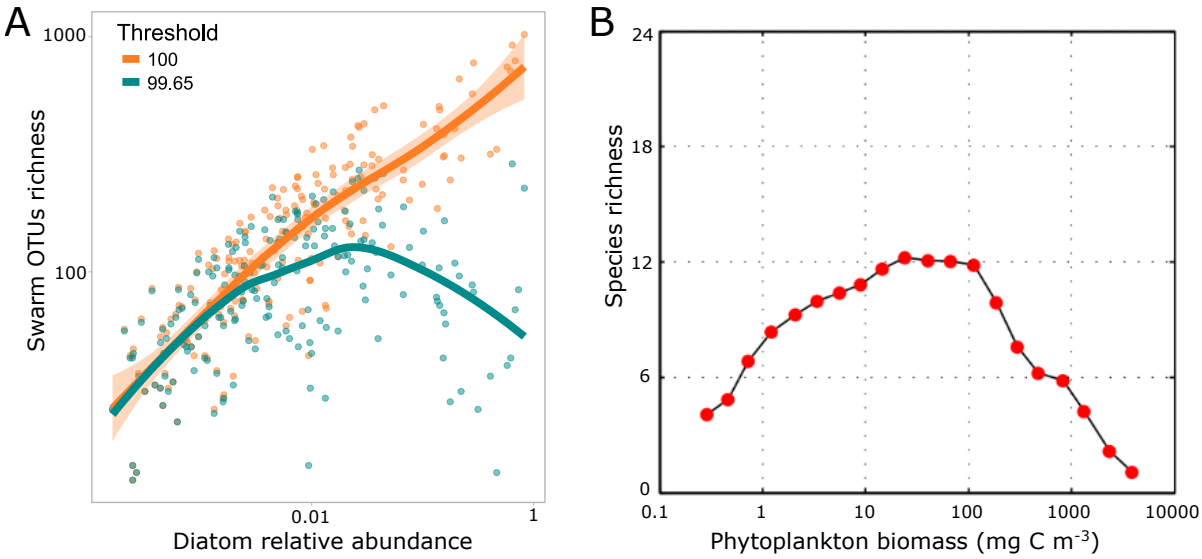

Fig. 2.7: Relationship between the Swarm metabarcoding diatom richness at two filtering thresholds (unfiltered = 100, orange; filtered = 99.65, cyan) and the relative abundance of diatoms in the sample computed as the sum of diatom Swarm OTU abundances over the whole Swarm OTUs abundances (panel A). In panel B) the figure from Vallina et al. (2014) showing the global productivity–diversity relationship (PDR) curve using equally spaced log10 bins of biomass. Noteworthy, the discard in absolute values between the scale of species richness deduced by this analysis (panel A) and the one obtained by Vallina et al. (panel B, Vallina et al., 2014a) is suggestive of the different filtering approaches of the two methods. Clearly, the approach applied by Vallina et al. takes into account only the very small, highly abundant subset of the actual species.

that the filtered richness is fitting this curve suggests that this is valid also for diatoms. It remains to be understood what is the nature of the discarded OTUs, whose abundance is very low but whose number is actually very large (e.g., in *Tara* station 173_SRF the discarded 0.35% of abundance corresponded to 566 OTUs, and in *Tara* station 205_SRF to 704 OTUs). In the literature there is also the hypothesis that productivity (resources availability) can increase richness (Mittelbach et al., 2001) and indeed in Vallina et al. (2014) diatoms have a richness that increases with primary production and only for very high values of production (i.e., only at the peak of a bloom) there is a drop. Thus, the linear increase for the unfiltered case is, in fact, not incompatible with current theories while indeed it refers to a very different dynamical scenario.

2.4.3 What is being filtered by the filtering process?

This filtering was applied to concatenate and reconcile the information retrieved from morphology-based identifications and metabarcodes. One fundamental difference between these two datasets is the definition of unit. The first approach aims to identify species through a process entirely based on their morphology, while the second is based on sequence similarity of the V9 subunit. Thus, the units identified by the metabarcode cannot be called species but rather operational taxonomic units (OTUs). OTUs are usually built to aggregate reads sharing 97% similarity, motivating by the expectation that those units may correspond approximately to the concept of species. However, the correspondence between OTUs and species is prone to error for three main causes:

1. Some species have genes more than 97% similar, resulting in OTU aggregating different species (Ratnasingham and Hebert, 2013);

2. Within the same species there may be paralogs <97%, providing more OTUs for a single species (Ratnasingham and Hebert, 2013);

3. Some OTUs may be spurious due to artefacts (e.g., read errors, chimeras; Brown et al., 2015).

Even though the Swarm clustering method stands on a number of single nucleotide differences (i.e., insertions, deletions or substitution) and not directly on measures of read similarity, the logical discourse remains the same. Whereas for the first point filtering the metabarcode cannot improve the correspondence between the number of OTUs and species, there is the possibility that the filtering procedure applied may correct errors induced by the other two causes listed above. The filtered OTUs thus result from artefacts and/or from very closely related species or strain within the same species.

How can we distinguish if the OTUs filtered are artifacts or evolutionary close species to the ones present?

The fact that I observed a linear relationship between diatom relative abundance and OTUs richness from the unfiltered dataset suggests that these filtered out OTUs may be linked to the amount of material sequenced. If we start by the hypothesis that saturation is equally reached in all the samples this information is supportive of filtered OTUs to be artifacts. The more diatom genomic material is sequenced the more artifacts are expected.

To deeper investigate the nature of the filtered OTUs, I analysed the phylogenetic diversity (PD) of samples before and after the filtering procedure. The rationale here is to observe how the phylogenetic diversity is affected by the filtering pipeline. A different behaviour from a linear relationship between the amount of OTUs excluded and the loss of phylogenetic diversity would be representative of one of two cases: either i) the filtered OTUs are equally spread across the phylogenetic tree and most of them are evolutionary close to more abundant OTUs or ii) the filtering cut off whole branches of the phylogenetic tree because rare OTUs are not randomly distributed (Fig. 2.8). I found the unfiltered dataset showing big peaks of PD in polar regions while the other regions have around one third of the PD found at high latitudes. Comparing the different scales between the two maps (Fig. 2.9) it is evident the strong loss of diversity induced by the filtering process.

The variation between the two maps is shown in Figure 2.10 as the delta percentage of PD. There is a strong latitudinal gradient of the impact of filtering on the PD as well. Polar stations are the most affected ones with a loss of phylogenetic diversity close to 100%. Comparing thus the loss of PD to the percentage of OTUs retained by the filtering (Fig. 2.11) I found a linear relationship between the two. The more OTUs are excluded the most is affected the PD of the same sampling station in a linear way. This may suggest

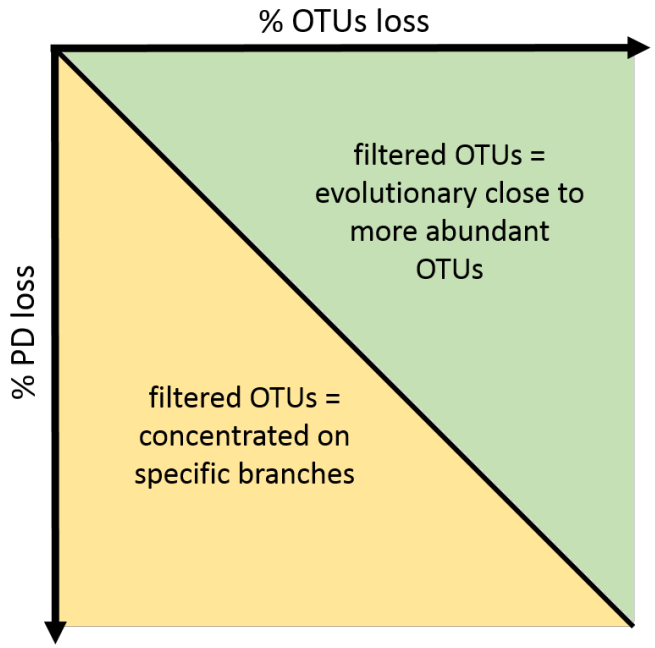

Fig. 2.8: Conceptual schema of the filtered OTUs distribution on the phylogenetic tree according to the relationship between the loss in PD and the percentage of filtered OTUs.

a balance between the two situations represented in Fig. 2.8. That is that filtered OTUs are likely strains or sub population evolutionary close to more abundant strains retained by the filtering process but widely spread across the phylogenetic tree.

To summarize, this preliminary investigation regarding the nature of these filtered OTUs suggests them to be artefacts, following the hypothesis that the more material is sequenced the higher is the probability to create artefacts. However, a second hypothesis is that in polar regions, which are also the more enriched of diatoms in terms of abundance, are areas characterized by a very large 'rare biosphere' of diatoms. This large reservoir of rare taxa would serve as pool of ecologically redundant species ready to become more abundant on the community whereas their optimal environmental condition would occur (Caron and Countway, 2009). This hypothesis seems to be more plausible given the phylogenetic diversity patterns. The polar stations are indeed characterized by a large phylogenetic loss induced by the filtering (Fig. 2.10), suggesting the loss of entire branches of the corresponding phylogenetic

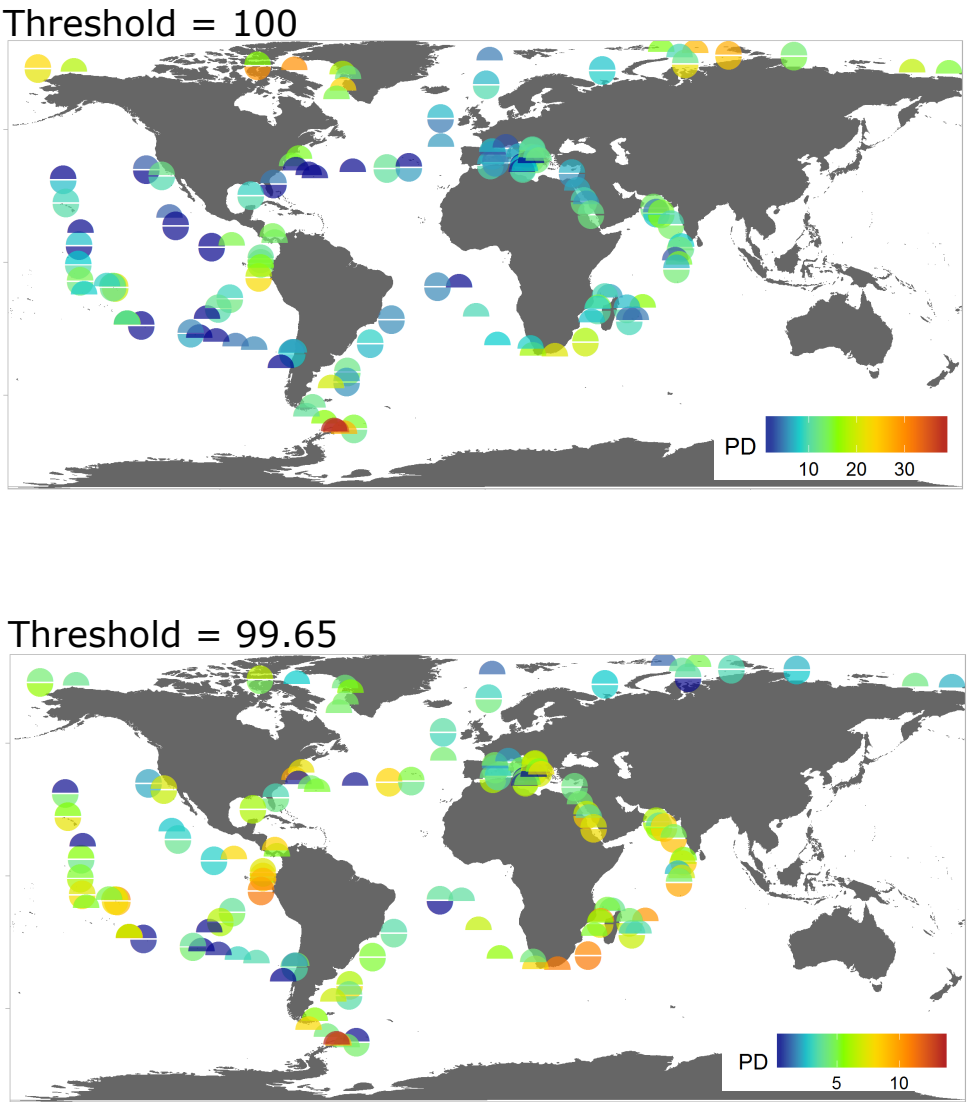

Fig. 2.9: Map of phylogenetic diversity across the sampling station based on unfiltered (t=100) and filtered (t=99.65) metabarcoding datasets at 20-180 μm.

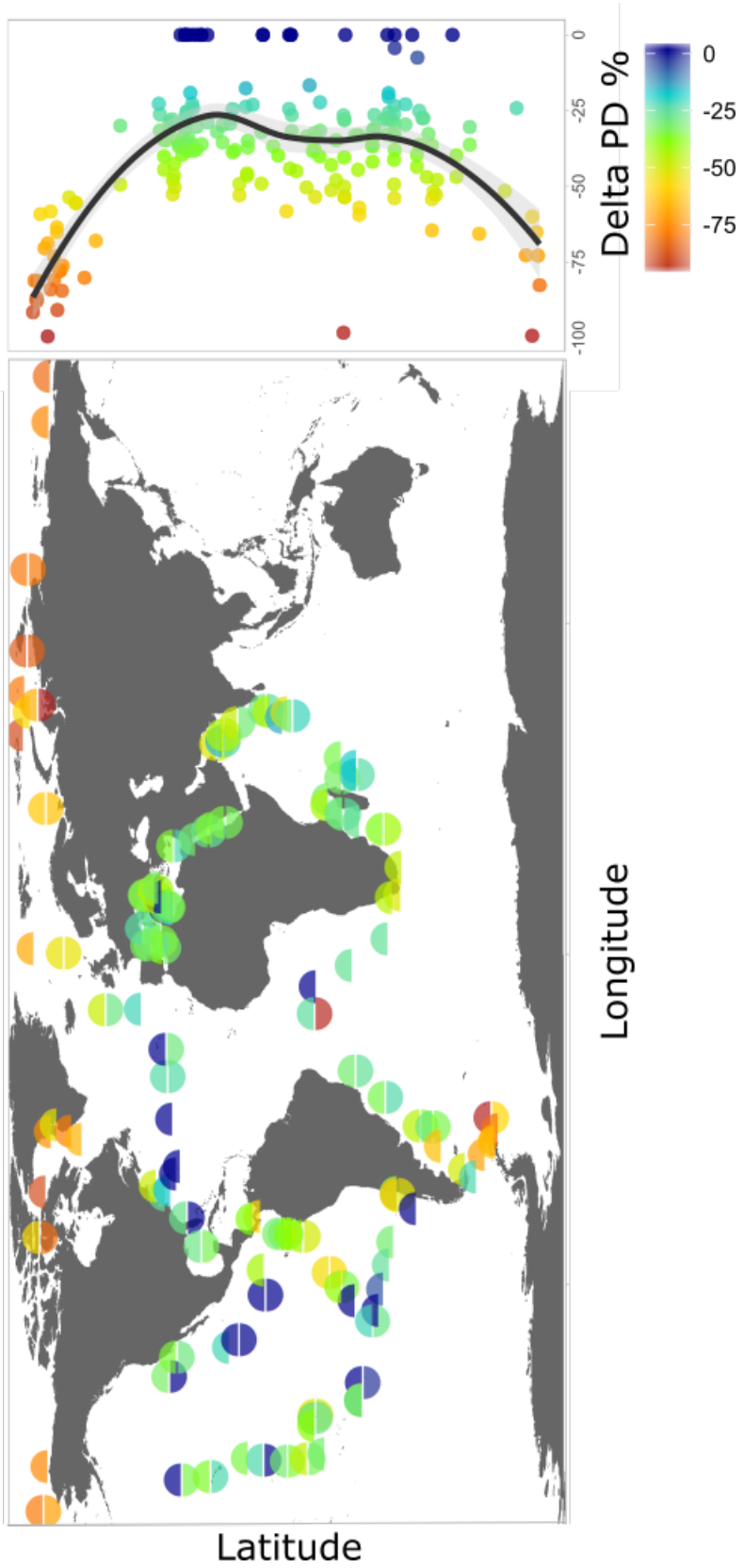

Fig. 2.10: Map of the delta of the PD between the filtered (T= 99.65) and the unfiltered diatom richness expressed as percentage over the unfiltered diatom richness. On the right a scatter plot of the same variable in function of the latitude with a linear curve fitted on the point distribution. Only negative delta are here shown, a delta PD % equal to zero means thus the absence of variation between the two PD or an increase of PD after the filtering process.

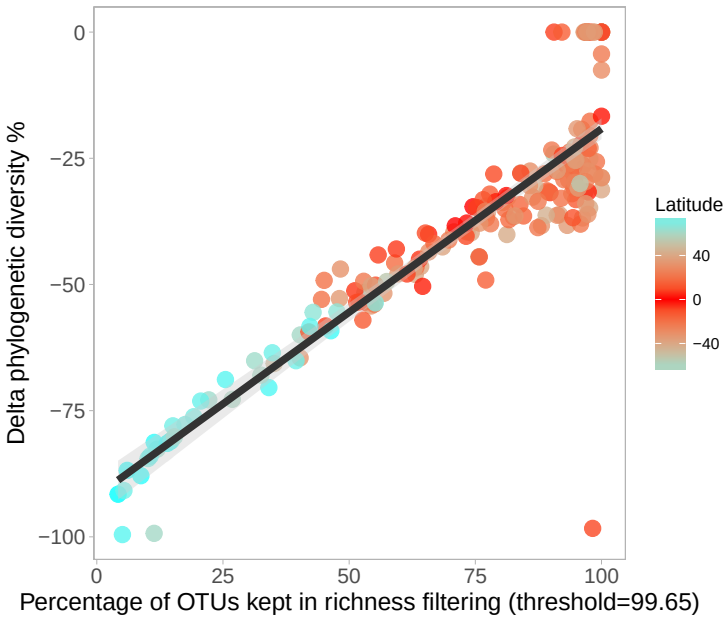

Fig. 2.11: Scatter plot of the percentage delta PD between the filtered (T= 99.65) and the unfiltered diatom richness and the percentage of OTUs kept in the filtering process over the unfiltered dataset. Each point corresponds to a sampling station and its color correspond to the latitude coordinate of the same station.

tree: branches incredibly rare but eventually ready to gain again an important role within the community.

2.4.4 Environmental and ecological drivers of diatom taxonomic richness

To characterize the environmental drivers of diatom richness I computed a BRT model of this same information. Nine prediction variables have been implemented for this exercise: a set of nutrients availabilities (iron, ammonium, nitrate, nitrite, silica), two hydrodynamics variables (gradient SST, lyapunov exponent), temperature and chlorophyll α as a proxy of the trophic system status. The contribution of these variables to the model highlights the almost equal role of flow-related variables, temperature and silica followed by a lower contribution of the other nutrients taken into account (Fig. 2.12). Noteworthy, temperature anticorrelates with nitrate and silica, and for this reason it is not really possible to state the effective direct role of temperature by itself.

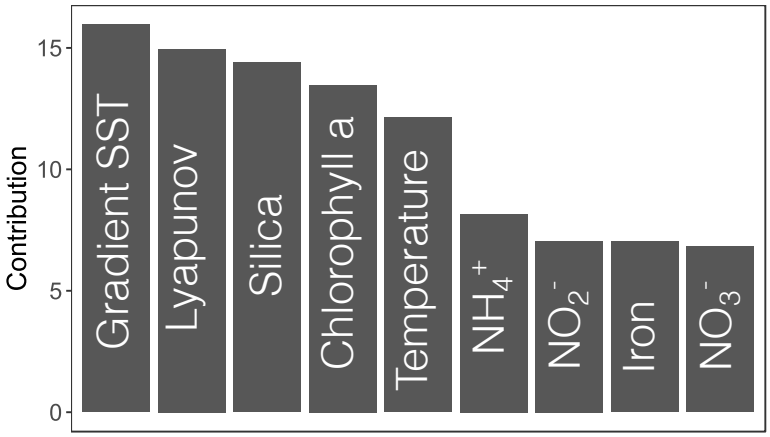

Fig. 2.12: Contribution of predictors defining diatom richness within the BRT model.

The model reaches very high predictive abilities, with a Pearson ρ correlation of 0.75 (p-value $< 2.2e^{-16}$) between the observed and predicted richness (Fig. 2.13). This is particularly surprising and interesting given that most of the variables refer to the *mean* state of the ocean.

The strongest limit of this model is in predicting the observed richness hotspot, and thus in sensing the processes behind the formation of the same. These peculiar stations diatom richness are indeed deeply underestimated by the model. This suggests either that there is not a unique global relationship between environmental variables and richness or that other factors are actually important in these extreme cases. Excluding these sampling stations from the Pearson correlation between the two information results in an increased ρ equal to 0.85 (p-value $< 2.2e^{-16}$), exhibiting the very high prediction ability of the model and thus its understanding of the linear and non-linear processes behind diatom richness.

The ecological hypothesis behind this chapter is that different processes play a role in defining diatom richness: according to the specific hydrographical structure, nutrient availability and hydrodynamics of the sites the dominating processes differ. To understand where lateral transfers, rather than nutrient availability are more informative in predicting diatom richness within the BRT

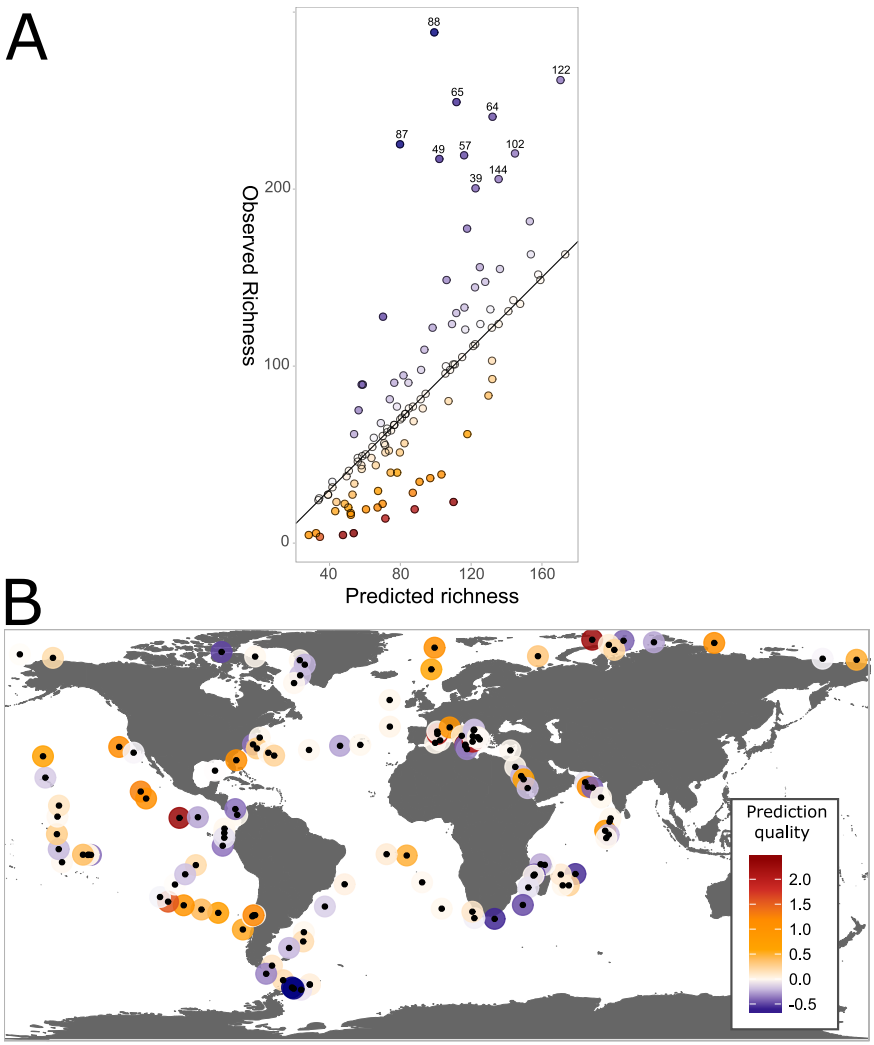

Fig. 2.13: In panel A the scatter plot of the richness as it has been observed from the filtered (t=99.65) metabarcode Swarm and as it has been modeled by the BRT model. Each point corresponds hence to a sampling station: they are labeled with the sampling station number where the observed richness is higher than 200. The color of the points corresponds to the model prediction quality expressed as the difference between the diatom richness predicted by the model and the diatom richness observed by the filtered metaB, normalized over the same observed diatom richness. In panel B the map of the prediction quality per sample. Positive values indicate stations where the model overestimated diatom richness while negative values indicate stations where the model underestimated it.

model I run a sensitivity exercise, running a model per predictor variable, excluding in each model the variable itself. Comparing the prediction quality of the original model to the partial model in each site give us an information of the informative power of each variable per site. Generally, the exclusion of one predictor rather than one other leads to very similar consequences on the prediction quality variation (Fig. 2.14). The fact that I obtain similar degrees of prediction variations excluding different variables is expected by the similar contribution that all the parameters play in the model (Fig. 2.12). However, the fact that negative and positive effects are very similarly distributed across the samples and variables would suggest that locally no process is particularly more important than one other. To deeply investigate this matter I projected the predictor effects on a heatmap, ordered according to the Euclidean distances between rows and columns (Fig. 2.15). Looking at this plot it is clear that while the above observation is true for most of the sampling stations, a cluster of stations actually have strong different answers excluding N sources, gradient SST, temperature and chlorophyll α compared to the other parameters. All these parameters have similar responses in prediction quality, clustered together with chlorophyll α, temperature and gradient SST. By contrast, silica and iron availability and Lyapunov exponent have slightly different quality effects on the prediction quality. The set of stations depicting a different importance of this second set of variables is widespread across the ocean, including a large variety of conditions. This may suggest the presence of an unaccounted process by the variables taken into account.

2.4.5 Diatom richness hotspot dynamics

The BRT modeling resulted in a very fine predictive tool to investigate the role of the different variables in shaping diatom richness with only one exception. Indeed, this machine learning approach failed to model the hotspots of richness, underestimating them. From the several processes proposed by the literature as main driver of phytoplankton diversity I found the two main

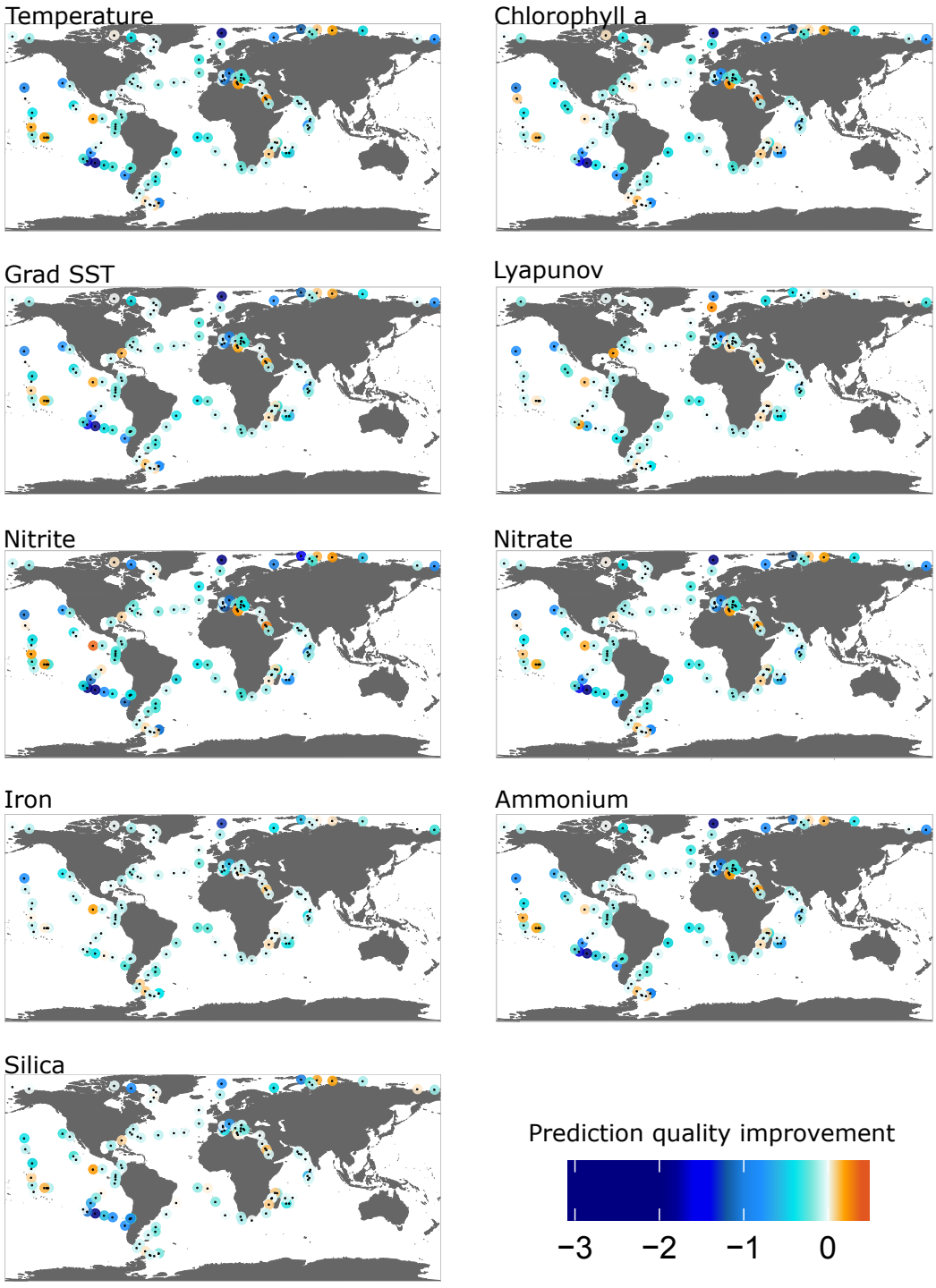

Fig. 2.14: Sensitivity exercise run per predictor variable. Each map shows the estimation of importance of the variable for each station. This information is measured by the improvement of prediction quality excluding one predictor variable from the model. Positive values indicate that in that station the exclusion of the environmental variable from the model actually improved the prediction ability of the machine learning model. Negative values indicate where the exclusion of the parameter lead to worse prediction than the original model potentiality. This means that negative values locate where the variable is important. Values are expressed in % over the observed diatom richness (see chapter 2.3.6).

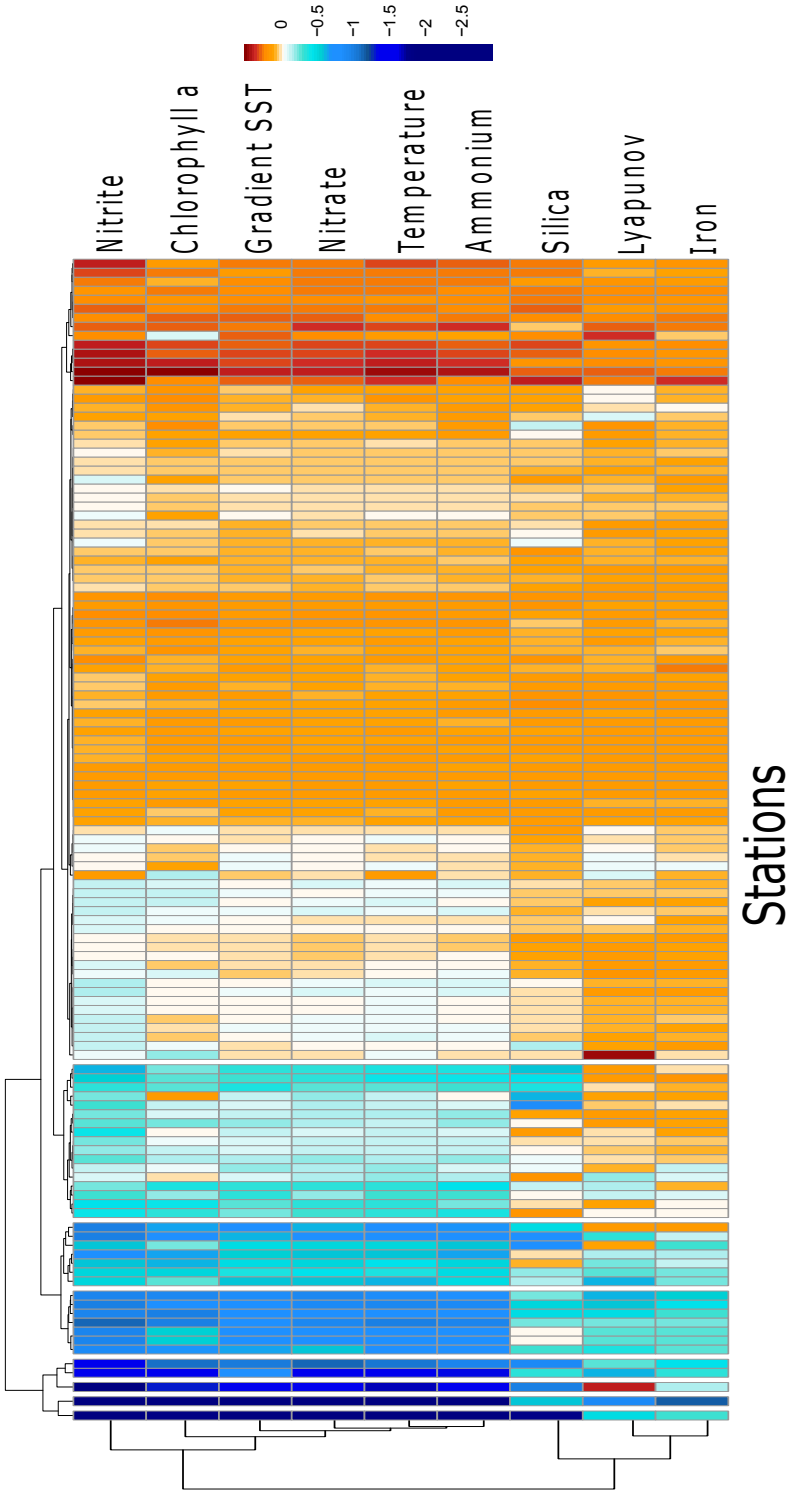

Fig. 2.15: Heatmap of the quality prediction improvement by predictor variable and stations. Rows and column are clustered through the complete method based on Euclidean distances between the vectors.

drivers to hotspots to be nutrient availability and water hydrodynamics processes. The role of the time variability (e.g., seasonality, mesoscale variability) is also considered (e.g., to explain the maxima in the Atlantic tropical regions, see Soccodato et al., 2016). The first is based on a direct bottom-up process: the high availability of nutrients allow the co-existence of a larger quantity of species (Dutkiewicz et al., 2009). By contrast, the second produces high hotspot of richness indirectly: water dynamics allow different communities to be mixed of overlapped in specific water structures (e.g., fronts, eddies), the presence of more communities on the same location results in a hotspot (Barton et al., 2010; Lévy et al., 2014). Indeed, there are other possible explanations, such as the time variability mentioned above or the presence of strong biotic interactions (e.g., symbiosis with bacteria Amin et al., 2015), but I do not have access to the information needed to test their importance, unfortunately.

The BRT model was not able to model the different processes behind the hotspots. One possible reason behind this inability of this approach lies in its search for global patterns or patterns completely distinguished by a variable through a decision tree. Applying BRT I can hardly understand the mechanisms behind each local value. In order to study the hotspots partially overlooked by this latter approach I complemented this analysis to a second one. To locate which processes dominate the other in the several diatom richness hotspots detected by the *Tara* Oceans expedition I run a particular machine learning technique, belonging to the neural network approach: the self organized maps (SOM). These maps cluster stations together on the basis of their environmental similarities. A model for each cluster of station is implemented and these latter are ordered in the space according to the diatom richness observed in the same stations. The models have been set on two set of parameters according to the scientific question. The first include a set of hydrodynamics variables while the second is a set of nutrient availabilities.

The resulting SOM detected a number of 4 clusters describing 4 different environmental situation leading to hotspot of diatom richness (Fig. 2.16). Cluster *blue* gathers together all the higher hotspots of richness (richness > 217 Fig. 2.16B), which are characterized by a relative importance of both hydrodynamics and nutrient parameters (Fig. 2.16A). These stations are widespread in the Antarctic, in the Indian and in the tropical Pacific Oceans. Cluster *yellow* is strongly characterized by iron availability: it groups together the hotspots observed in the Mediterranean Sea stations and the tropical Pacific stations. The Mediterranean sea shows typically enhanced phytoplankton diversity by iron deposition from the Sahara dust, while the Marquese island depicts a bloom provoked by sudden iron availability (Caputi et al., Submitted). Hydrodynamics variables dominates only the cluster *green*, which actually is represented only by a *Tara* Ocean station located in the Gulf Stream. This region of energetic ocean circulation was expected to result in diversity hotspot because of the strong lateral dispersal (Barton et al., 2010). Finally cluster *pink* is located in areas enriched in nutrient but also characterized by strong hydrodynamics, representing an overlap between the enhancements of diversity by the two processes.

2.5 Conclusions

Two main scientific questions have been addressed in this chapter. The first one is essentially methodological: How can I reconcile a diatom taxonomic richness measure from omic and morphology- based information?, whereas the second one is purely ecological: What are the environmental drivers of this diversity?. I firstly designed a way to combine diatom richness measures obtained by two different approaches and consequently investigated the processes behind this ecological index patterns through statistical means.

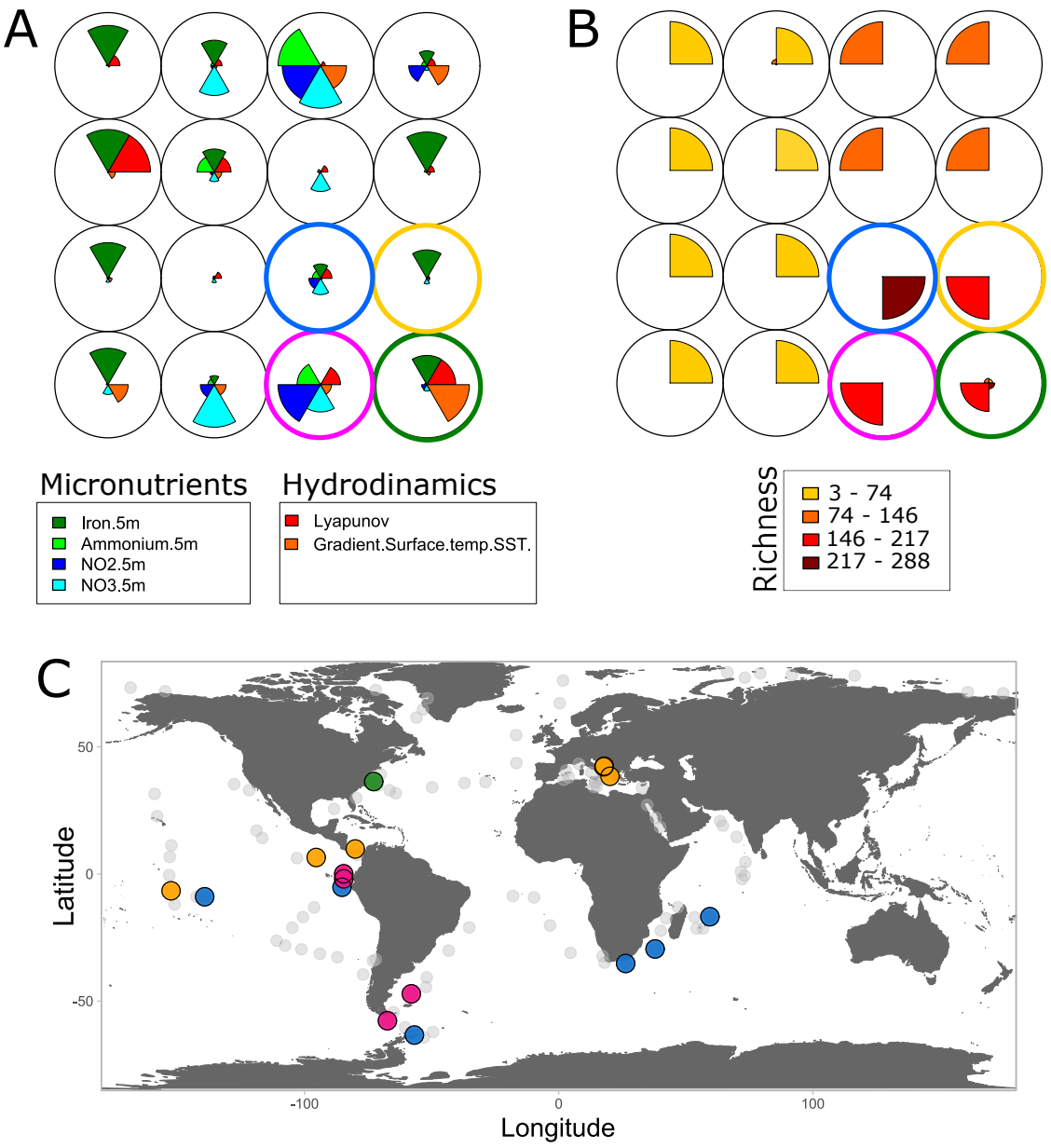

Fig. 2.16: Extension of SOMs to multiple data layers: panel A and B correspond to the SOM of the layer of environmental variables and diatom richness respectively. In panel C the stations clustered in four nodes of the above SOMs are mapped over the global ocean. The four clusters are identified by the color of the contour of each node: *pink*, *blue*, *green* and *yellow*, corresponding to the same colors applied in panel C.

Previous studies have already compared metabarcoding and morphology based information for phytoplankton (e.g., Abad et al., 2016) or even specifically for diatoms (e.g., Zimmermann et al., 2015; Malviya et al., 2016). The comparison of these two kind of data is not to be given for granted. From the two examples cited above, it is expected for metabarcode to measure a widely larger number (up to two folds) of OTUs compared to the number of species morphology-based identified. The distribution of abundances across the OTUs is different from the abundances estimated by microscopy counts: not always the OTUs found to be abundant correspond to the species observed as abundant through the morphology-based identifications (Stoeck et al., 2014; Hirai et al., 2015; Massana et al., 2015; Sun et al., 2015; Abad et al., 2016). The quantitative power of metabarcoding has still to be better investigated but the reasons behind these incongruities between the different measured relative abundances may be explained by: i) technical bias during the DNA extraction due to different performances according to the organism type or the development stage (Roh et al., 2006), or ii) during the PCR amplification step (Gonzalez et al., 2012) which may favour abundant taxa in the amplification process, or iii) due to the copy number variation associated with rDNA (Kembel et al., 2012). Moreover, another limit of metabarcoding lies in the taxonomic identification of the OTUs which is strictly linked to the reference catalogue available. Focusing on diversity measures allow to describe the communities avoiding OTU taxonomic identification. In particular, given the strong limits in OTU quantitative measures I found it more appropriate to use as diversity index a measure which would not take into account the abundances of its units: the richness.

The pipeline herein designed focuses on filtering the metabarcoding according to the relative abundances of OTUs to obtain a correspondence between the filtered metabarcode richness and the morphology-based richness. It could be extremely significant to include also the metagenomic and metatranscriptomic information to derive the melting point of the four different

sources. Unfortunately, we are still unable to obtain such type of information from these two latter kinds of data. Compared to the unfiltered metabarcode, the metabarcode I obtained through the optimal filtering showed global richness patterns close to what was expected for phytoplankton (Barton et al., 2010). As Malviya et al. (2016) were the first one to provide global insights on diatom distribution, this study is the first attempt to describe global diatom richness.

But what was actually excluded through this filtering process? The sequences excluded by the filtering process may be artefacts (Brown et al., 2015) and/or, more probably, the so called 'rare biosphere', composed by ecologically redundant units represented by rare sequences evolutionary close to unfiltered more abundant sequences. This second option is totally expected if we consider the fact that we are shaping the metabarcode to be more similar to the microscopy- based observation. Morphology-based observations are often limited to the genus level, and differently from metabarcode they most certainly don't have the resolution to identify strains or cryptic species (Zimmermann et al., 2015). Moreover, it is supported by the phylogenetic analysis I ran. Future efforts should be made to study the identity of these rare species, both taxonomically, looking at their phylogenetic relationship to the more abundant ones, and functionally, to investigate the ecological role they may play in the community if they were more abundant. Understanding the taxonomic and ecological reservoir of the community was not the direct aim of this chapter, but it is surely a fundamental field of ecological research which can not be foreseen while working with richness estimations. Further studies can now be developed on this matter thanks to the insights on rare species provided by metabarcoding approaches.

Interestingly, the impact of the filtering process was not equally weighted across the regions: the sampling stations which endured the exclusion of the larger percentage of OTUs were all located at the poles. Further investigation

are needed to explain the reason of this higher abundance of rare species in the poles compared to the other stations, but among the possible interpretations I reported here two plausible hypothesis of the processes behind this specific distribution.

The first hypothesis is that the poles are sites of faster speciation rates compared to other regions and this is the origin of the larger number of rare OTUs at high latitudes. The possibility of faster speciation rates at the poles rather than at the tropics has been recently hypothesized (Rabosky et al., 2018) and it is based on the concept that high rates of speciation occur in areas with low surface temperatures and high endemism. A possible driver of this higher speciation rates may be a process of metabolic isolation driven by the extreme environmental situation, particularly in terms of light and temperature conditions. Also, in extreme environments stress is higher and this usually increase the mutational rates (Galhardo et al., 2007). Moreover, selection pressure may be reduced in these regions as the hard conditions of these latter make the diatom temporal window of occurrence rather narrow, resulting in short vegetative season and reducing thus risks such as grazing.

A second hypothesis able to justify the higher number of OTUs filtered at the poles compared to the other station is a process of accumulation in these areas. Independently by the speciation rates distribution, OTUs move from the region of origin to the poles, the solely place where they linger. This means not only that the oceanic circulation allows to a very high number of species to reach this region but also that the environmental conditions of the poles allow to the species itself to survive in vegetative or resting phases. Phenomena of accumulation have already been observed in phytoplankton for other regions and explained by combination of specific ocean hydrodynamics (Villar et al., 2015). This observation in any case could motivate a series of dedicated studies and new observations. Finer-scale molecular studies focused on mutations such as SNPs on this same omic data may allow insights on population dynamics

and thus on the mechanisms beneath the formation of these hotspots. What limits a robust interpretation is the non linear relationship between the *in situ* abundance and the obtained sequences. As the amount of sequences obtained from the samples slightly differs, this may represent a bias particularly for the detection of rare species.

Once I had obtained the reconciled richness measure for all the *Tara* Oceans stations and assessed the effect of the filtering procedure, I addressed the biological question of this chapter: What are the processes supporting and shaping diatom richness? The implemented BRT modeling successfully modeled the distribution of diatom richness across the *Tara* Oceans sampling stations. Using as predictor nutrients, hydrodynamics, temperature and chlorophyll α concentration was enough to explain the large majority of the observed data. Astonishingly, the developed model is able to predict diatom richness with an accuracy of 0.8 excluding hotspots local regions. This tool results hence in a very powerful predictor of diatom richness answer to climate change. The proper prediction by global biogeochemical climate change models of the environmental variables used as explicative variables in the BRT model could allow deep insights on the global distribution of diatom richness over time.

Peculiarly, I found that diatom richness can be enhanced by low iron concentrations, suggestive that the rare iron fertilization and eddies can likely support not only the growth of low iron adapted species but also species which prefer high iron concentration and are able to survive also at very low concentrations. Noteworthy, diatom richness is enhanced as well by low availabilities of ammonium and nitrate, indicating that in condition of strong competition over nutrient resources the pool of species present is richer, ready to allow more complex community dynamics over seasons and eventually sudden nutrient inputs or climate changes. Interestingly, the higher hotspots of diversity were not well predicted by the model. My hypothesis is that hotspots can results from different conditions: for high availability

of nutrients (Dutkiewicz et al., 2009) or for water dynamics overlapping communities from different regions (Barton et al., 2010; Lévy et al., 2014). The supervised SOM applied to two sets of environmental data to explain diatom richness captured four different environmental conditions which lead to diatom hotspots. According to my results, hotspots can be induced by iron availability, which is often a limiting nutrient (Coale et al., 1996), by lateral transport dynamics (Lévy et al., 2014; Lévy et al., 2015), but also by the overlap of the two processes, where the higher diatom richness values are observed. The integration of other key variables of these processes within the BRT model could improve it to the point of correctly predict these hotspots too. Further efforts should be made on the study of hydrodynamical dynamics impact on diatoms and on the proper way to measure it.

3

The identification of putative functional units of N transporter gene families

3.1 Summary and main achievements

- Chapter 3 to 5 present a new perspective on the assessment of functional diversity from metatranscriptomic and metagenomic data;

- Through this chapter I assessed diatom functional units from metatranscriptomic and metagenomic datasets. They are defined as the phylogenetic clades of high-affinity ammonium (*AMT1*) and nitrate (*NRT2*) transporters gene families. This original working framework is based on the hypothesis that diatoms evolved different genes to solve this function to linger in different conditions;

- I described the evolutionary scenarios of these two gene families as a complex series of duplication and gene loss events;

- I established the taxonomic composition of the designed functional units and assessed the independence of these latter from a taxonomic bias;

- I detected the limits of the metagenomic dataset to work at this fine scale due to lack of saturation;

- Finally, I provided a broad overview on the evolutionary clades, highlighting their low ubiquity and the large numeric range of unigenes composing them.

3.2 Introduction

Species cannot be considered equal in their impact on ecosystem functioning (Mouchet et al., 2010) but this disparity is completely neglected by taxonomic diversity. The concept of functional diversity arose thus to better describe the ecosystem together with its stability and functioning. This diversity is not supposed to be a substitute of taxonomic diversity but rather a different descriptor of the community. The differences between the two diversities are manifold. While in taxonomic diversity metrics the units are just different taxa, functional diversity aims to group species according to their role in the ecosystem. This creates a big discrepancy between the two as the number of different functional units present may strongly differ from the number of different taxonomic units present. Of course this comparison strongly depends on the resolution of the taxonomic ranking and of the functional definition, but anyway, conceptually, they are describing two different properties of the system. Among the advantages of this new diversity, functional classification has allowed i) to evaluate similar systems characterized by different taxonomic compositions, ii) to improve habitat classification, particularly of habitats characterized by large number of species of difficult identification taxa (Salmaso et al., 2015), iii) to improve water quality assessment (B-Béres et al., 2016) and iv) to study ecosystem functioning variations (Török et al., 2016).

The real difficulty of functional diversity lies in the functional classification of taxa. Profusely different functional classifications have been proposed, based on completely different criteria. One option has been to aggregate elements (i.e., taxa) sharing the same structural and/or functional features:

species that behave so similarly to be described by a single parametrization of functional traits. Among these, we can include morphological criteria (e.g., cell size, form), physiological criteria (e.g., N fixation, silica users), life strategies characteristic (e.g., single cells, colonial life), temporal appearance (e.g., spring bloomer), distributional characteristics (e.g., polar species) but also taxonomic units at low classification levels (Hood et al., 2006; Irwin et al., 2006; Finkel et al., 2010). An alternative solution has been to classify functional classes according to the co-occurrence of taxa on a representative spatial or temporal dataset.

Throughout the literature, this latter approach was applied by the first functional classification of phytoplankton, which goes back to Reynolds (Reynolds, 1980). Focusing on co-occurrence of phytoplankton species he defined 14 functional groups as the clusters of taxa co-occurring together across the time-series. The same author (Reynolds, 1988) proposed a different grouping key, classifying phytoplankton taxa in three phytoplankton groups according to the susceptibility to stress, disturbance and limiting resource utilization of the single species. The assignation of phytoplankton species to functional groups had several follow-ups. Typically, phytoplankton groupings are based on a combination of the previously cited criteria, mainly cell size, higher phylogenetic grouping and biogeochemical roles like, for example, silicifying diatoms, N fixing and non-N fixing cyanobacteria, mixotrophic dinoflagellates and calcifying coccolithophorids. But it is exactly the grouping of different taxa that is the main challenge in defining functional diversity because there is no unique correct choice in the selection of traits. Traits should be chosen according to the scientific question of the study, taking into account how the ecosystem process of interest works and which organisms and traits are most influenced by this same process (Petchey and Gaston, 2006). Moreover, the number and type of traits selected have a strong impact on the observed diversity, affecting the number of units within the system. Again, the

number of traits taken into account has to be selected in order to describe the specific function of interest (Petchey and Gaston, 2006).

In marine studies, diatoms are thus generally classified as a unique functional group, united by their biogeochemical role as silicate users. However diatoms are one of the most diversified taxonomic groups, which reflects very different physiologies and lifestyles (chapter 1). Therefore, freshwater studies have further developed diatom functional descriptions, classifying the latter in different functional units. Passy (2007) published the nowadays most applied functional classification of freshwater diatoms, calling them diatom guilds. Three guilds were established, and taxa were classified within those according to their use of resources and disturbance tolerance. Rimet and Bouchex (2012) updated this classification by adding a fourth guild for planktonic diatoms, later assessed as ecologically fundamental for these systems (B-Béres et al., 2017), and further re-assigned several taxa to different guilds. Also concerning freshwater studies, traits such as cell size or biovolume were used to study the relationship with the ecosystem environmental descriptor. These single traits were found to be strongly correlated to nutrient uptake and efficiency (Tapolczai et al., 2016), trophic levels (Berthon et al., 2011), physical disturbances (Tapolczai et al., 2016) and salinity (Kókai et al., 2015). But both the use of guilds or single traits proved not to be robust enough to model and predict ecosystem processes (B-Béres et al., 2016). Even though freshwater studies have already advanced work on functional classification within the diatoms during the last decade, marine research still has not addressed the problem and the strong intrinsic difference of the two ecosystems make the functional classification developed for freshwater systems completely inadequate for marine ones.

In this chapter I propose a new classification of diatom functional diversity based on a single trait: the resource utilization trait. Resource utilization traits are one of the key traits in describing temporal and spatial population

dynamics and competition among microbes, but they are also fundamental to predict biogeochemical impacts (Litchman et al., 2015a). In order to exploit the unprecedented massive meta-omics material produced by the *Tara* Oceans expedition I selected specific gene families that are representative of the resource utilization trait.

I chose to work with N metabolism related genes as nitrogen is one of the main limiting nutrient for phytoplankton. Nitrogen metabolism in diatoms needs to be tightly regulated to immediately react to the fast environmental changes that characterize the ocean. Among the diatom genes involved in the nitrogen metabolism, several have been proposed as potential markers to assess the response of cells to different stress, in particular nutrient starvation. Among these genes we list nitrate reductase (*NR*), glutamine synthetase (*GLNII*) and the ammonium and nitrate transporters (*AMT* and *NRT*; Allen et al., 2005; Song and Ward, 2007).

The nitrate reductase expression in diatoms has no circadian regulation, contrary to other phytoplankton, but it is instead induced or inhibited by the metabolic pools of nitrate and ammonium respectively (Brown et al., 2009). Indeed, its expression is down-regulated by NH_4^+ availability and upregulated in case of NO_3^- availability, to the point that its activity is correlated with external NO_3^- concentration (Berges and Harrison, 1995). Indeed, in case of large availability of nitrogen or phosphate its transcription has been found largely induced (Alexander et al., 2015). But not only replete NO_3^- conditions induce its expression, as deplete conditions for the same nutrient induce it as well (Mccarthy et al., 2017). The activity of this gene is strictly inversely linked to temperature between 12°C and 25°C (Gao et al., 1993; Lomas and Glibert, 1999).

Glutamine synthetase has been proposed as marker for its fundamental and delicate role as key enzyme in the coupling of N and C metabolism of

the cell (Parker and Armbrust, 2005). This enzyme catalyzes the production of glutamine from the reduced NH_4^+ and the glutamate generated by the GOGAT within the plastid. The activity of this enzyme is commonly used as measure of productivity through *in situ* measurement (Rees et al., 1995). Studies on the environmental and cellular regulators of GS in diatoms are quite recent. Takabayashi et al. (2005) found GSII to be induced by NH_4^+ availability from reduction of NO_2^- and not by the NH_4^+ directly obtained by the environment (also Bender et al., 2012). For this reason GSII is a good marker of new production. Its transcription has been found to be induced by high light and low temperature (12°C) as well as in cells grown on nitrate rather than ammonium (Parker and Armbrust, 2005; Bender et al., 2012). Nevertheless, more recently Alipanah et al. (2015) did not find any regulation of this gene in nitrogen starvation conditions while Alexander et al. (2015) found a downregulation in starvation from phosphates or nitrates.

Ammonium and nitrate transporters, like other N metabolism molecular markers, show differential expression in different environmental conditions. Several environmental cues may have a role in diatoms expression patterns, the most investigated one is nitrogen availability. Numerous authors investigated differential expression under N starvation conditions, in availability of NH_4^+ or NO_3^-. Species-specific responses have been obtained (Bender et al., 2014), with also different regulation of the multiple copies of the same gene within the same species (Tab. 5.1 and 5.2, see Rogato et al., 2015 and references therein; Alipanah et al., 2015). More generally, in phytoplankton NO_3^- transporters have been found induced by the presence of nitrate, while ammonium transporters are rather induced in condition of starvation due to absence or low availability of ammonium (Glibert et al., 2016 and references therein).

Within this context I chose as marker high-affinity N transporter genes, both for the ammonium and the nitrate source of N, because of the relevance

of their direct role in the cell-environment relationship. Within this metabolic pathway, the selected high affinity ammonium (NH_4^+ – *AMT1*) and nitrate (NO_3^- - *NRT2*) transporter families are of peculiar interest as it is assumed that they enable diatoms to handle N fluctuations in the global oceans (Rogato et al., 2015). Two distinct gene families constitute *NRT*s: *NRT1* and *NRT2*, representing respectively low and high affinity nitrate transporters, which show no overall sequences similarity (Wittgenstein et al., 2014). *AMT*s too are composed of two gene families: *AMT1* and *AMT2*, representing respectively high and low affinity ammonium transporter, having a common distant evolutionary origin (Wittgenstein et al., 2014). In land plants *NRT1*, *NRT2*, *AMT1* and *AMT2* gene families all display a monophyletic status; however a complex scenario of lineage specific gene duplication and loss emerged early in land plant evolution, suggesting that evolutionary divergences have mirrored functionally distinct groups (Wittgenstein et al., 2014).

The pipeline I propose herein allows the exploitation of the different genes belonging to the same gene families through phylogenetic analysis. Indeed, for each function ('ammonium uptake' or 'nitrate uptake') the cell owns a set of genes encoding proteins that perform the same role. Following the working hypothesis that evolutionarily related genes have similar modulations I propose as putative functional units of diatoms the phylogenetic clades of genes as described by these two gene families.

3.3 Material and Methods

3.3.1 Data

Sequence searches

Diatom *AMT1* and *NRT2* sequences from a previous report (Rogato et al., 2015) were used as reference datasets to search for putative diatom orthologues in the *Tara* Oceans metatranscriptome database (Alberti et al., 2017; Carradec et al., 2018). Searches were performed using CLADE, which designed a domain architecture for each reference sequence. Hidden Markov model (HMMs) profile corresponding to the detected domains (available in Pfam database) were saved into an initial pHMM database. This database was enriched by adding three new pHMMs: one built from the entire set of reference sequences, and two others built exclusively from diatoms and non diatoms reference sequences. The pHMM database was then used to scan the six translations of *Tara* Oceans metatranscriptome dataset in protein sequences, according to the six possible reading frames of the sequences. For that, we used HMMer version 3.1 (with -cut_ga option) and detected more than 6,000 *Tara* sequences as the referred transporters. Among them a considerable amount of false positive is expected, even though the pHMM database was composed only by true reference transporter sequences. To reduce false positives, a second run of CLADE over the 6,000 *Tara* sequences (all six frame translations) was performed by the most probable domain architecture for each sequence. We analysed these domain architectures and only sequences containing at least one domains of pHMM database were considered (10% of the initial set of sequences).

To select the most probable translation, DAMA (Domain Annotation by a Multi-objective Approach, Bernardes et al., 2015) was modified to consider

the six frames. We consider as putative transporter only the most probable frame translation that present the same domain architecture of the reference sequences. The putative transporters were then split into diatom and non diatom species. For that, we checked the taxon agreement of pHMM model (first analyses), and of CLADE results (second analyses). Only sequences with diatom taxon on both models were considered to be true diatoms transporters. 529 putative diatoms *AMT1* (*Tara*-di-*AMT1*) and 471 putative diatoms *NRT2* (*Tara*-di-*NRT2*) sequences were retrieved from the search.

The analysis described in this section were performed by Dr. Fabio Vieira Rocha from the Ecole Normale Supérieure of Paris.

Phylogenetic analysis

The same set of known *AMT1* and *NRT2* transcripts from Rogato and co-authors (Rogato et al., 2015) was used as query against 99 Diatoms transcriptomes at the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP - Keeling et al., 2014) using an in-house developed BLASTp search. From this search a total number of 45 *AMT1* (MMETPS-di-*AMT1*) and 51 *NRT2* (MMETPS-di-*NRT2*) transcripts were obtained. The putative *Tara*-di-*AMT1* and *Tara*-di-*NRT2* sequences resulted by CLADE were thus 6-frame translated into amino acid sequences through the EMBOSS Transeq web site (`http://www.ebi.ac.uk/Tools/st/emboss_transeq/`; Li et al., 2015). A catalogue of diatoms sequences was built from the *Tara*-di-*AMT1* and *Tara*-di-*NRT2* amino-acid sequences and the MMETSP sequences. This catalogue together with appropriate non-diatoms sequences were aligned with Muscle (Edgar, 2004) using basic parameters. Two trimmed alignment were obtained. The di-*AMT1* alignment (Supplementary File 5) is composed of 282 sequences, of which 162 are *Tara*-di-*AMT1*. The alignment consists of 137 AA positions (including gaps). The trimmed di-*NRT2* alignment (Supplementary File 6) includes 259

AA sequences (166 *Tara*-di-*NRT2*) and consists of 108 positions (including gaps). Consensus sequences for the *Tara*-di-*AMT1* and the *Tara*-di-*NRT2* alignments were graphically represented using sequence logo (Schneider and Stephens, 1990) at the Web Logo website (`http://weblogo.berkeley.edu/logo.cgi`; Crooks et al., 2004). Phylogenetic analyses were then performed using i) Neighbour Joining method (Saitou and Nei, 1987), ii) approximately-maximum-likelihood method (aML – Yang, 1994; Anisimova and Gascuel, 2006) and iii) Bayesian Inference (BI, Mau et al., 1999) approaches. The NJ phylogeny was inferred in MEGA 7 (Kumar et al., 2016) using 10,000 bootstrap replicates. The evolutionary distances were computed using the JTT matrix-based method (Jones et al., 1992). To infer the aML phylogenetic relationships we used the FastTree2 software (Price et al., 2010). The BI analysis was conducted using MRBAYES v3.2 (Ronquist et al., 2012). Trees were sampled every 1,000 generations for six million generations, and the first 25% of all the trees sampled were discarded as burn-in. From the phylogenies were manually defined 11 clades for the di-*AMT1* (Supplementary File 7) and 12 clades for the di-*NRT2* (Supplementary File 8).

The analysis described in this section were performed by Dr. Luigi Caputi from Stazione Zoologica Anton Dohrn and Dr. Maurizio Chiurazzi from CNR.

Taxonomic annotation

The taxonomic identification of the *Tara* Oceans di-*AMT1* and di-*NRT2* unigenes retained by the phylogenetic analysis was done through a blast search against all the proteins presents in the MMETSP dataset (Keeling et al., 2014) using a local BLAST portal.

The best blast-hit was retained and compared to the taxonomic identification of sequences performed by CLADE (2.3.1). Mismatches between the

two identifications were estimated at genus levels. The sequences annotated as *Extubocellulus* and *Odontella* through the MMETSP database produced all mismatches as CLADE lacks the probabilistic model for these two genus. Other than these two genus a number of 47 sequences over the 588 taken into account for both gene families produced mismatches (8%). These latter are probably due to the different phylogenetic tree of diatoms used by MMETSP (SILVA/PR2) and CLADE (NCBI/uniprot). I chose therefore to consider the MMETSP-based taxonomic identification where there are mismatches. Taxonomic annotation for di-*AMT1* and di-*NRT2* can be found respectively in Supplementary File 7 and 8.

The analysis described in this section were performed by Dr. Luigi Caputi from Stazione Zoologica Anton Dohrn and Dr. Fabio Vieira Rocha from the Ecole Normale Supérieure of Paris.

Transporter abundances and normalization

The abundance of the unigenes selected by the phylogenetic analysis was then extracted by the metatranscriptomic and metagenomic dataset of *Tara* Oceans (Carradec et al., 2018). Occurrences values are computed as the fraction of the number of reads mapped per kb of unigene covered with reads per the total abundance of reads annotated to diatoms in the sample. The sum of occurrences for all the unigenes for a given sample is equal to 1. The dataset of di-*AMT1* and di-*NRT2* occurrences within the metatranscriptome dataset can be found respectively in Supplementary File 9 and 10 while in the metagenomic dataset the data refers to Supplementary File 11 and 12. The total number of reads sequenced per sample is available as well in Supplementary File 13, for the metatranscriptome and in Supplementary File 14 for the metagenome. A summary of the four datasets can be found in Tab. 3.1

The data was extracted by the cited public datasets by Dr. Eric Pelletier from Genoscope in Paris.

3.3.2 Data mining

The presence-absence of each clade for both families is defined by the presence in the metatranscriptome database or the metagenome database of *Tara* Oceans. A clades is considered present in a sampling site if the sum of the mRNA abundance of the unigenes belonging to the clade is higher than 0 in at least one of the four size-classes sampled (0.8-5 μm; 5-20 μm, 20-180 μm; 180-2,000 μm). A clade mRNA abundance is computed per each size class and per sample as the sum of the abundances of unigenes belonging to the same clade.

3.3.3 Saturation analysis

Species accumulation curves (SAC) were drawn randomly permuting samples for 1000 permutations and estimating the mean SAC together with its standard deviation (Gotelli and Colwell, 2001). Samples were weighed over the sampling effort expressed as the number of reads sequenced per sample. The average of species richness per sample is thus calculated from linear interpolation of single random permutations. The curves were plotted for both gene families and for both metagenomic and metatranscriptomic datasets. Chao estimation of total richness was then estimated for the same data. These analysis were performed through the functions *specaccum* and *specpool* of the R package *vegan* (Oksanen et al., 2017).

Tab. 3.1: Summary table of the content of the Supplementary Files 9 to 12, that is the metagenomic and metatranscriptomic occurrences of the unigenes associated to the gene families di-*AMT1* and di-*NRT2*. For every size-fraction and sampling depth it is reported the the minimum, maximum and median occurrence of unigenes together with the number of unigenes detected (Count).

Gene	Size	Depth	Min occ.	Max occ.	Median occ.	Count
			MetaTranscriptome			
AMT1	0.8-5	DCM	7.54E+07	2.92E+11	1.41E+09	466
AMT1	0.8-5	SRF	2.66E+07	4.57E+11	8.78E+08	858
AMT1	180-2000	DCM	9.53E+07	4.16E+11	2.76E+09	191
AMT1	180-2000	SRF	3.93E+06	2.54E+11	2.01E+09	393
AMT1	20-180	DCM	1.72E+07	7.61E+11	1.97E+09	938
AMT1	20-180	SRF	6.72E+06	4.25E+11	1.51E+09	1514
AMT1	5-20	DCM	1.23E+07	4.78E+11	1.66E+09	931
AMT1	5-20	SRF	6.50E+06	3.74E+11	1.34E+09	1659
NRT2	0.8-5	DCM	7.67E+07	2.75E+11	1.46E+09	238
NRT2	0.8-5	SRF	3.14E+07	1.43E+12	8.69E+08	696
NRT2	180-2000	DCM	1.51E+08	1.19E+11	3.65E+09	76
NRT2	180-2000	SRF	4.11E+06	1.57E+11	2.57E+09	262
NRT2	20-180	DCM	2.63E+07	2.53E+11	1.52E+09	663
NRT2	20-180	SRF	1.20E+07	4.27E+11	1.55E+09	1160
NRT2	5-20	DCM	1.66E+07	6.33E+11	1.18E+09	550
NRT2	5-20	SRF	7.91E+06	3.02E+11	8.90E+08	1296
			MetaGenome			
AMT1	0.8-5	DCM	2.67E+08	1.21E+11	3.98E+09	238
AMT1	0.8-5	SRF	1.12E+08	8.37E+10	4.31E+09	696
AMT1	180-2000	DCM	1.84E+09	1.14E+11	7.92E+09	76
AMT1	180-2000	SRF	1.52E+08	9.28E+10	6.15E+09	262
AMT1	20-180	DCM	3.22E+08	1.22E+11	5.13E+09	663
AMT1	20-180	SRF	2.00E+08	1.34E+11	3.99E+09	1160
AMT1	5-20	DCM	3.18E+08	1.76E+11	4.18E+09	550
AMT1	5-20	SRF	7.29E+07	5.17E+10	1.89E+09	1296
NRT2	0.8-5	DCM	2.37E+08	1.83E+11	5.12E+09	238
NRT2	0.8-5	SRF	1.12E+08	1.39E+11	3.94E+09	696
NRT2	180-2000	DCM	1.16E+09	9.77E+10	1.05E+10	76
NRT2	180-2000	SRF	1.32E+08	1.38E+11	6.88E+09	262
NRT2	20-180	DCM	2.86E+08	1.34E+11	5.27E+09	663
NRT2	20-180	SRF	2.18E+08	2.67E+11	5.33E+09	1160
NRT2	5-20	DCM	2.57E+08	6.61E+11	5.82E+09	550
NRT2	5-20	SRF	1.09E+08	1.36E+11	2.77E+09	1296

3.3.4 Unigenes

Comparison between the metagenome and the metatranscriptome datasets is based over the presence-absence datasets computed over both databases. Errors were counted as number of times each unigene was observed in a station in the metatranscriptome but not in the corresponding sample of the metagenome.

Ubiquity of unigenes is expressed as number of stations each gene is present in the metatranscriptome over the total number of stations taken into account. Consequently, median ubiquity of clades is expressed as the median of the ubiquity of the single unigenes constituting the clade.

3.4 Results and Discussion

3.4.1 Phylogenetic trees

To derive a diatom functional description from metatranscriptomic and metagenomic data I selected a trait of interest, such as N uptake, and investigated the different genes able to perform this function. Working under the hypothesis that evolutionary close genes have similar functions, a deep evolutionary analysis of two N uptake gene families, *AMT1* and *NRT2*, has been performed. Following this logic, clustering genes according to their phylogenetic relationship would allow to discriminate genes according to their functional role. The assessment of phylogenetic clades is hence a mean in this framework to define different functional units. With this purpose, for the two gene families a phylogenetic tree was inferred and rooted over non-diatom N transporters sequences (Fig. 3.1). Both of them suggest that the diatom part of the tree is monophyletic, with no sign of lateral gene transfer. While previ-

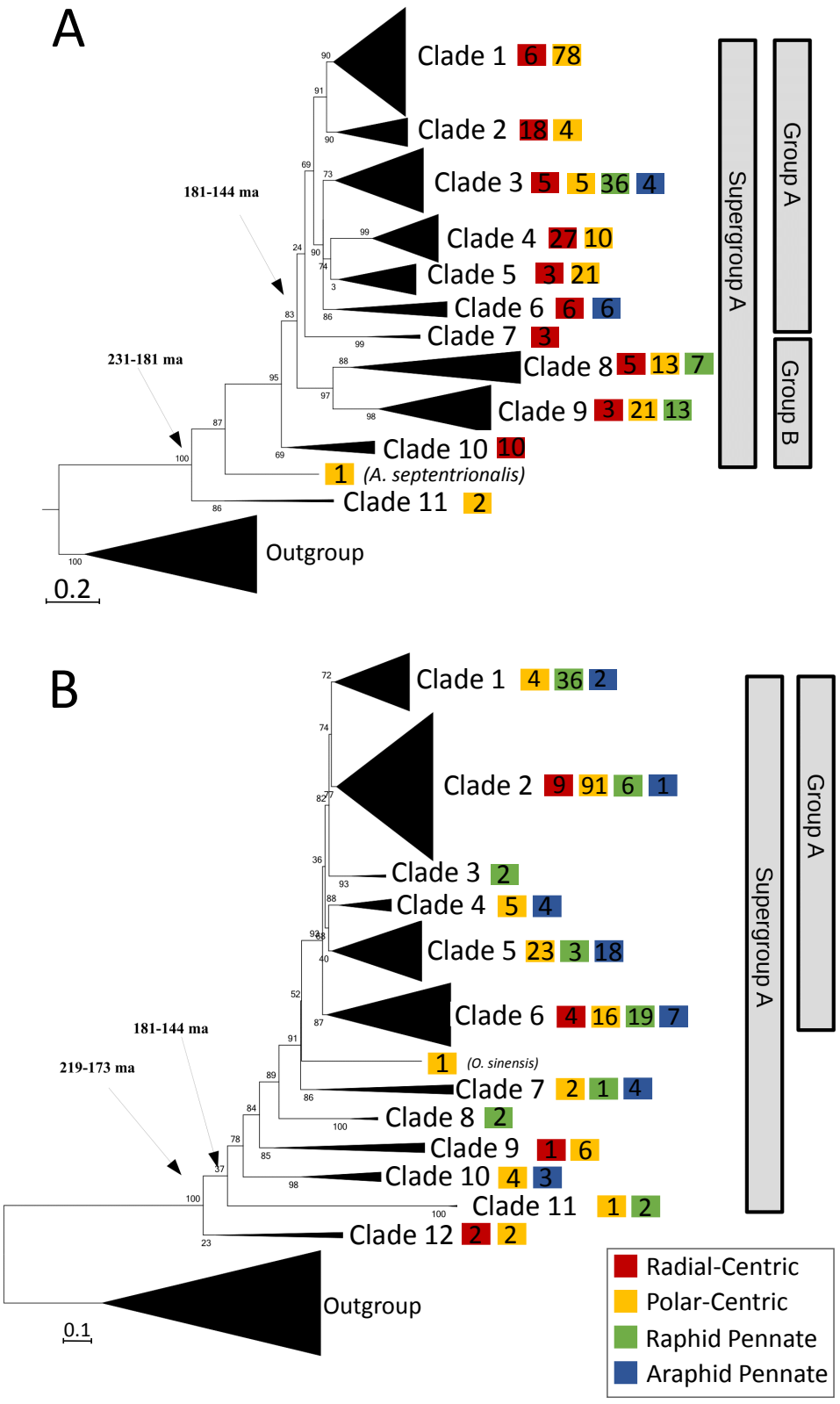

Fig. 3.1: Phylogenetic relationships of A) diatom di-*AMT1* and B) di-*NRT2*. Branch thickness indicates the statistical support for the clades. Thick lines indicate bootstrap values >50% (NJ), Shimodaira-Hasegawa (SH) >50% (aML) and a posterior probability >0.5 (BI). Medium thickness lines indicates bootstrap values >50% (aML). Branches not satisfying any of these parameters were collapsed. Clades have been collapsed and annotated according to the taxonomic assignment of the transcripts to the major diatom groups. The taxonomic assignment is designated by a colored square, within which is written the number of sequences annotated.

ous reports investigated the phylogeny of AMTs and NRTs in diatoms using a limited repertoire of sequences from published genomes (McDonald et al., 2010; Wittgenstein et al., 2014; Rogato et al., 2015) herein it is proposed a new phylogeny including an unprecedented number of sequences.

Eleven clades have been identified within the di-*AMT1* phylogenetic tree (Fig. 3.1A). Within the latter tree topology there is a dichotomy displaying clade 11 basal to the supergroup A (clades 2-10). Within the supergroup A all clades included at least a Radial-centric-basal-Coscinodiscophyceae (RC) sequence. Clade 10 is basal to the rest of supergroup A and the fact that it includes only RC di-*AMT1* suggests the ancestry of this clade. In supergroup A clades owned by Polar-centric-Mediophyceae and/or by Raphid Pennates emerged early in group B (clades 8-9) and in the whole group A (clades 1–6). Pennate diatoms do not own any sequence on clades 1 or 2 and Araphid Pennate diatoms have representatives only in clades 6 and 3, as consequence of likely group specific gene loss events. Thus, evolutionary reconstruction of di-*AMT1* relationships within diatoms exhibits a complex scenario. Results are suggestive of various round of an ancestral repertoire of di-*AMT1* which expanded from the Radial-centric-basal-Coscinodiscophyceae to the Polar-centric-Mediophyceae, undergoing specific gene duplications and losses in the Pennate diatoms.

The di-*NRT2* tree is very complex and not fully resolved, depicting diatoms sequences distributed across 12 clades (Fig. 3.1B). Clade 12 owned by Radial-centric-basal-Coscinodiscophyceae and by Polar-centric-Mediophyceae diatoms is basal to the supergroup A, built from clades 1-11, suggesting that the latter diverged around 219-173 ma (Medlin, 2016). The presence of Pennate diatoms in the basal clades of the supergroup A (Raphid Pennate in clades 11 and Araphid Pennate in clade 10) strongly suggests that the radiation of di-*NRT2*, followed by multiple rounds of gene loss and duplication, occurred prior to the evolutionary divergence of the two Pennate groups (181 – 144 ma

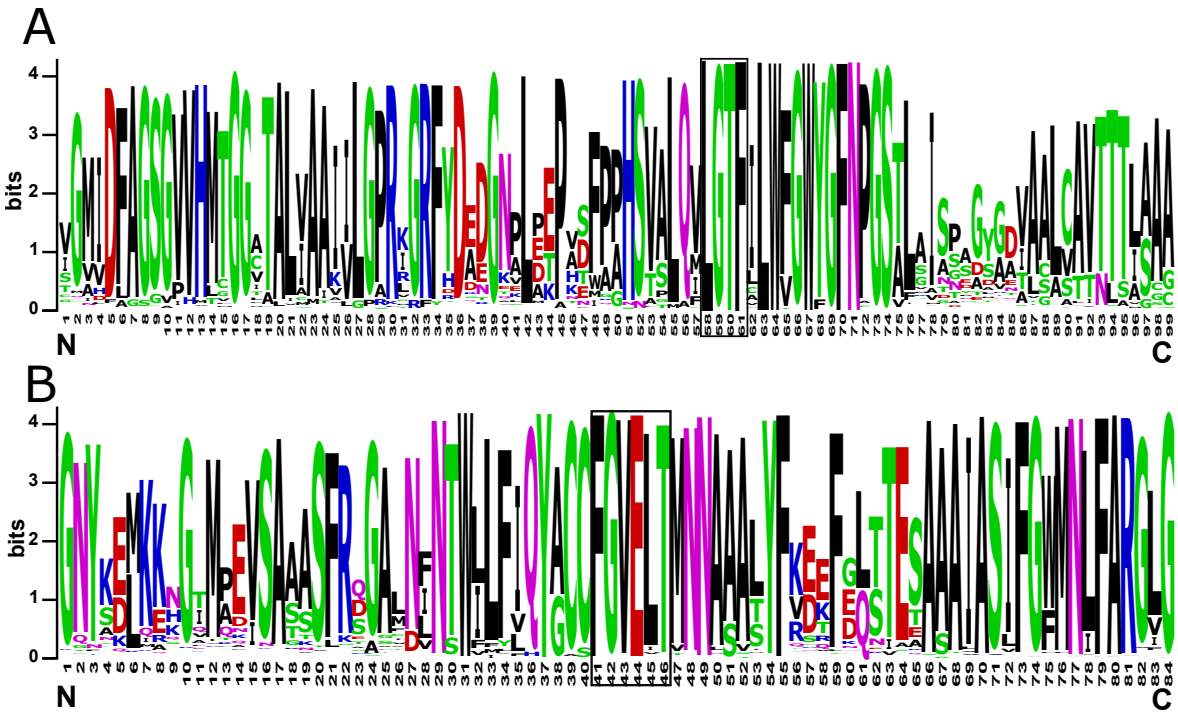

Fig. 3.2: Sequence logo consensus for di-AMT1 (A) and di-NRT2 (B) alignments. Conserved regions are framed in the logo and they corresponds to the 'LGTF' sequence in di-AMT1 (A) and 'FGVELT' sequence for di-NRT2 (B).

- Medlin, 2016). Only two clades within group A (namely clade 2 and clade 6) belong to all the main four diatoms groups, supporting the hypothesis of specific gene loss and duplication. Interestingly, two clades (clade 3 and 8) only includes unigenes assigned to Raphid Pennate, which may indicate recent functional diversification.

Abovementioned results indicate that diatom N transporters evolved through complex and not fully intelligible steps of duplications and losses, likely to respond to specific needs of functional specialization and diversification.

3.4.2 Conserved regions

To investigate the presence of conserved regions across this two gene families, and to compare it to what is known on terrestrial plants, alignment and logo analysis were performed across the retrieved *Tara* Oceans N trans-

porter genes. The trimmed NRT2 alignment included 281 putative di-NRT2 sequences. The total length of the alignment was 222 AA (including gaps). 22 sites were conserved among the whole dataset and 143 AA positions were parsimony informative. The di-AMT1 alignment included 307 sequences and had a total length of 202 AA position (with gaps). In the alignment, 23 AA positions were found to be conserved in sequences, and 139 out of 202 AA position were parsimony informative.

Concerning the conserved regions, two short amino acid stretches were found to be preserved along the great majority of the di-AMT1 and di-NRT2 alignments. A 4 AA motif (LGTF) was 100% conserved at position 58-61 of the di-AMT1 alignment (Fig. 3.2), the same signature is shared with land plants di-AMT1 sequences where they fall into the 6th TM domain. di-AMT1 proteins are predicted to display 11th TM domains with a typical N-out C-in topology (Rogato et al., 2015) and possess 13 out of the 14 conserved amino acid residues reported to be functionally significant for conducting ammonium through the pore region (Andrade and Einsle, 2007).

The FGVELT motif represents a signature for most of di-NRT2 proteins, located within the 7th TM domain (Fig. 3.2). Such a motif is only partially conserved in *Aspergillus nidulans* and dinoflagellates (four out of six), whereas two different signature motifs have been identified within TM5 and TM11 (Unkles et al., 2012; Dagenais Bellefeuille and Morse, 2016). The glutamate residue (E) within the FGVELT motif is conserved in all eukaryotic nitrate transporters and salt-bridges formation with other conserved amino acids have been hypothesized with the function of stabilization of protein conformation (Unkles et al., 2004). 12 TM domains have been predicted for most of the di-NRT2 proteins (Rogato et al., 2015) those central loop is significantly longer (35/38 aa) than in plant NRT2 proteins (21 aa: Kotur et al., 2016).

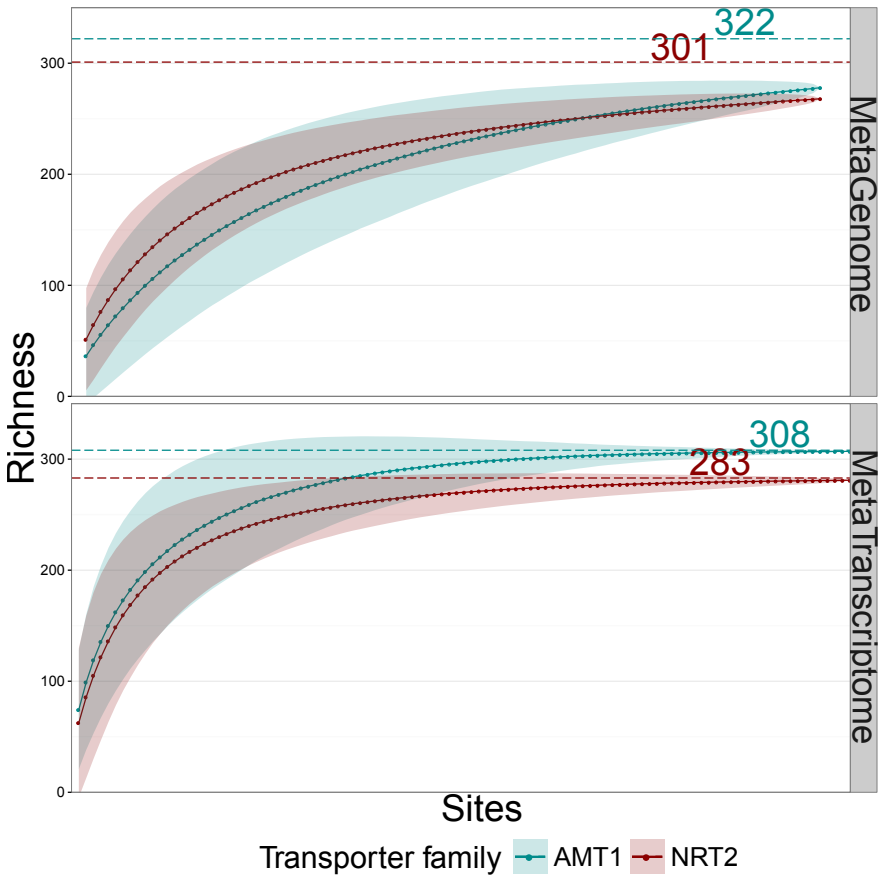

Fig. 3.3: Saturation curves of di-*AMT1* (blue) and di-*NRT2* (red) in the two omic datasets: metagenomic (upper panel) and metatranscriptomic (lower panel). Chao estimation of total richness are annotated over the SAC as horizontal lines, whose intercepts correspond to the estimation.

3.4.3 Sequences search

The number of sequences found in the *Tara* Oceans database assigned to diatoms according to the previous phylogenetic analysis counts 307 genes for di-*AMT1* and 281 di-*NRT2* genes belonging to diatoms. The number of diatoms species has several estimations, ranging from 4,700 (Malviya et al., 2016), to 30,000 (Guiry, 2012) and up to 100,000 species (Mann and Vanormelingen, 2013). From literature (Rogato et al., 2015) there are only four species of diatoms for those the N transporters genes of ammonium and nitrate were identified: *Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, *Fragilariopsis cylindrus* and *Pseudo-nitzschia multiseries*. For di-*AMT1* the number of gene copies in these species is of 7, 8, 8 and 5 while for di-*NRT2* were found 3, 6, 9 and 3 respectively. Using these (very limited) numbers of transporters and multiplying it for the estimated number of diatom species (see above), the total number of different transporters should be ranging from 23,500 to 800,000 di-*AMT1* and from 14,100 to 900,000 di-*NRT2*.

The numbers I obtain are far from these rough estimations and to investigate the number of unobserved transporters genes I run saturation analysis and total richness estimates (Fig. 3.3). Saturation curves are very promising in the metatranscriptome dataset, where both gene families seems to reach the plateau corresponding to the same Chao estimation of total richness. According to these first observations, within the limits of *Tara* Oceans sampled area, I collected close to the totality of expressed N transporter genes. The situation is fairly worse for the metagenomic dataset, where the plateau is not reached by neither of the two transporter families. There is a wide saturation problem in this dataset which makes it very problematic for further analysis. The limits of metagenomic datasets will be soon addressed and investigated in chapter 3.4.5.

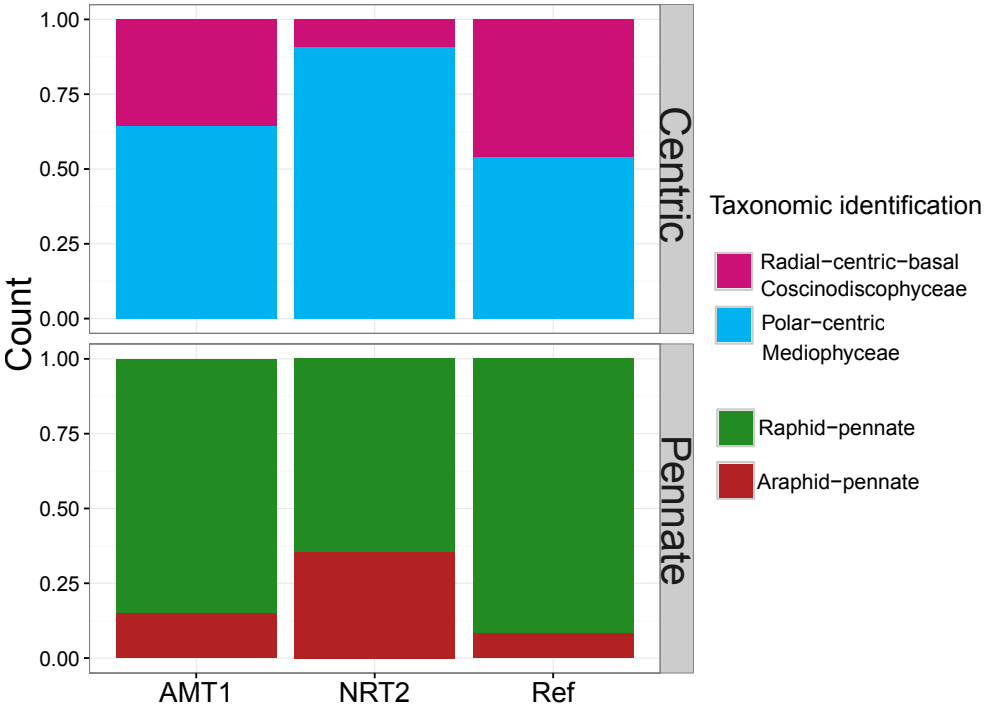

Fig. 3.4: Histogram of the taxonomic assignation of the sequences to major phylogenetic diatoms groups, compared to the number of diatoms species present in the database Algaebase for the same groups (1,300 Coscinodiscophyceae, 1,531 Mediophyceae, 11,434 Raphid Pennates and 1,040 Araphid Pennates).

3.4.4 Taxonomic identification

Of the 307 unigenes unambiguously assigned to diatom *AMT1* (di-*AMT1*), 50% was assigned to Polar-centric-Mediophyceae, 28% to Radial-centric-basal-Coscinodiscophyceae, 18% to Raphid Pennate and the remaining 4% to Araphid Pennate. The diatom *NRT2* (di-*NRT2*) dataset is also dominated by Polar-centric-Mediophyceae (55% of the total di-*NRT2*), while only the 6% of the unigenes were assigned to Radial-centric-basal-Coscinodiscophyceae. Overall, the 39% of the di-*NRT2* unigenes were assigned to Pennate diatoms (25% to Raphid Pennate and 14% to Araphid Pennate, respectively). If we compare the relative distribution of major taxonomic groups to the distribution of species ever observed by taxonomist (Algaebase) we obtain a fine correspondence between di-*AMT1* distribution and the references one but a high underestimation of Radial-centric-basal-Coscinodiscophyceae in di-*NRT2* (Fig. 3.4). We cannot compare pennate diatoms with the references as the sampling

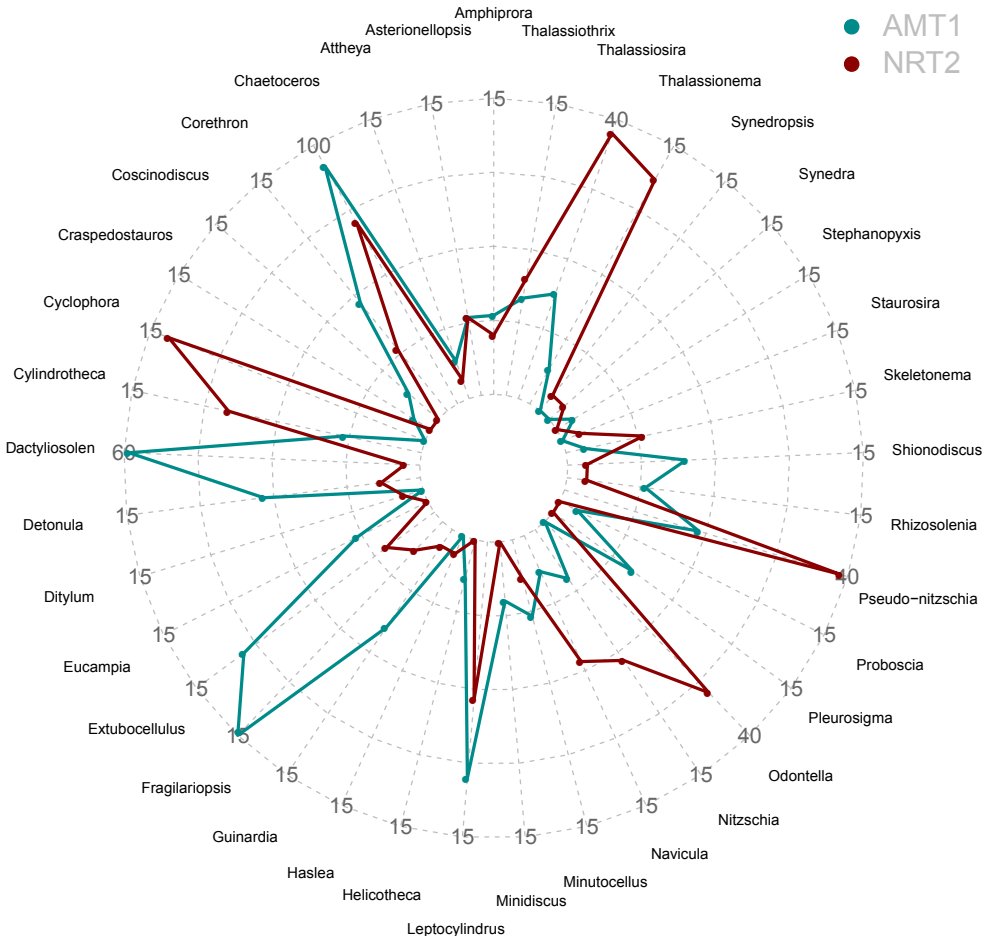

Fig. 3.5: Spider chart of the number of genes of di-*AMT1* (blue) and di-*NRT2* (red) assigned to the different genera. All the axis start at zero while the peripheral labels indicate the maximum values of the correspondent axis.

concerned only the water column and diatoms include many benthic forms, mostly belonging to the pennate group. Of course we have to keep in mind that references are biased towards the species living in the most accessible areas, but the same bias is present in the taxonomic identification of sequences, forcely linked to the available reference libraries.

If we look closer at the quantitative differences between the taxonomic identification of di-*AMT1* and di-*NRT2* we found very different representation of genes (Fig. 3.5). Malviya et al. (2016) found the intragenic diversity of diatoms on the metabarcode dataset of a subset of *Tara* Oceans stations to be particularly high in *Pseudo-nitzschia*, *Chaetoceros*, and *Thalassiosira* genera, while it was very low in *Corethron*, *Leptocylindrus*, *Minidiscus*, and *Planktoniella*. I found coherency for the higher differentiated genera, which are the same that have the higher number of transporters, but compared to that genera, most of the others are found with extremely low number of transporter genes.

It appears from the taxonomic identification of genera that the selected genes are a good representation of diatoms diversity, well distributed across all genera. For this reason even if saturation was not reached, I expect the sampled sequences to be representative of the whole diatom taxonomic group.

3.4.5 N transporter unigenes

As consequence of the SAC curves obtained for both gene families in the metagenomic dataset I deepened the investigation on the correspondences between the two datasets. I firstly looked at the differences between the presences of genes observed in the metagenome and the corresponding presences in the metatranscriptomes. Biologically, it is possible to observe a gene (metaG) but not its transcript (metaT) in the omics analysis of the same sample if its transcription is not active. However, it is biologically impossible for a gene that is transcripted (metaT) to not be found present in the metaG,

for obvious reasons. The presence of these errors is mostly due to lack of saturation of the sequencing, which may originate from low abundant genes highly expressed.

Tab. 3.2: Number of incongruences between metagenomic and metatranscriptomic datasets. Each case corresponds to occurrence of gene present in the MetaT but not in the corresponding MetaG sample. # of cases corresponds to the total number of occurrences; # of genes is the number of genes showing at least once this difference between the datasets; # of stations refers to the number of stations having at least one genes showing this difference.

	# cases	**# genes**	**# stations**
di-*AMT1*	3,490	299	106
di-*NRT2*	2,353	272	106

As resumed in Tab. 3.2 this error is widely spread across the genes and stations, but if we carefully observe the distribution of this error (Fig. 3.6 and 3.7) it would seem that this error is not linked to specific stations, covering all the *Tara* Oceans sampling. By contrast, it would appear that specific genes are harder to detect in the metagenome rather than in the metatranscriptome. This could suggest a specific modulation of a subset of species which would highly express N transporters in case of rare concentrations.

As few stations had a higher number of error occurrences I checked for the corresponding number of reads sequenced in the samples as well as the number of reads belonging to diatoms. It seems that stations enriched of metatranscriptome diatoms reads are the ones prone to error but this is not linked to the sequencing depth of single samples but rather to the quote of reads belonging to diatoms (Fig. 3.8 and 3.9). Samples quantitatively enriched in diatoms may have thus an higher number of false negative in the metagenome, but also an higher number of transporter sequences observed from both datasets, as for example surface stations 64, 80 and 84.

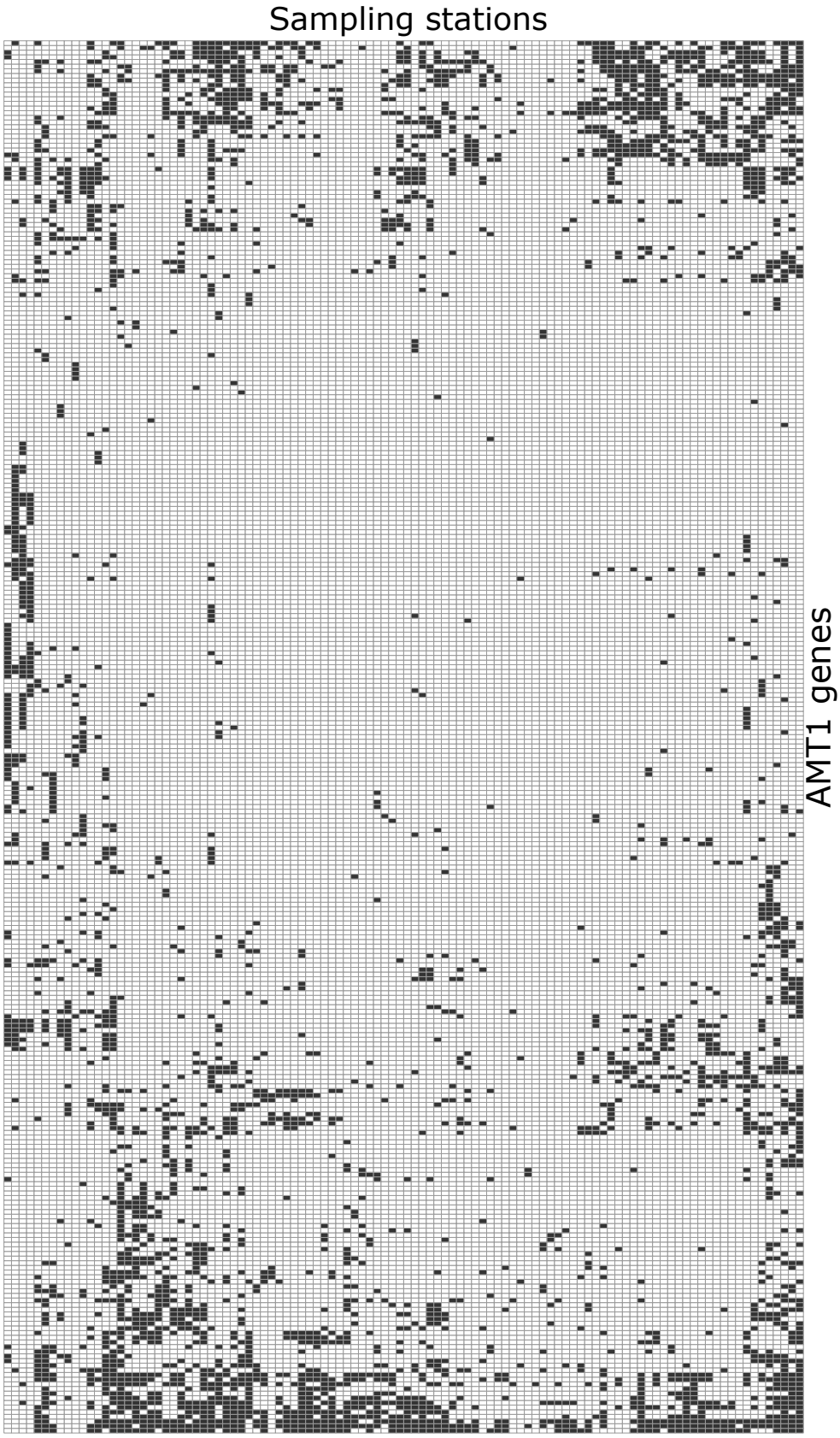

Fig. 3.6: Heatmap of saturation errors occurrences for di-*AMT1*. Grey cells correspond to the stations where the corresponding transporter gene shows this type of error.

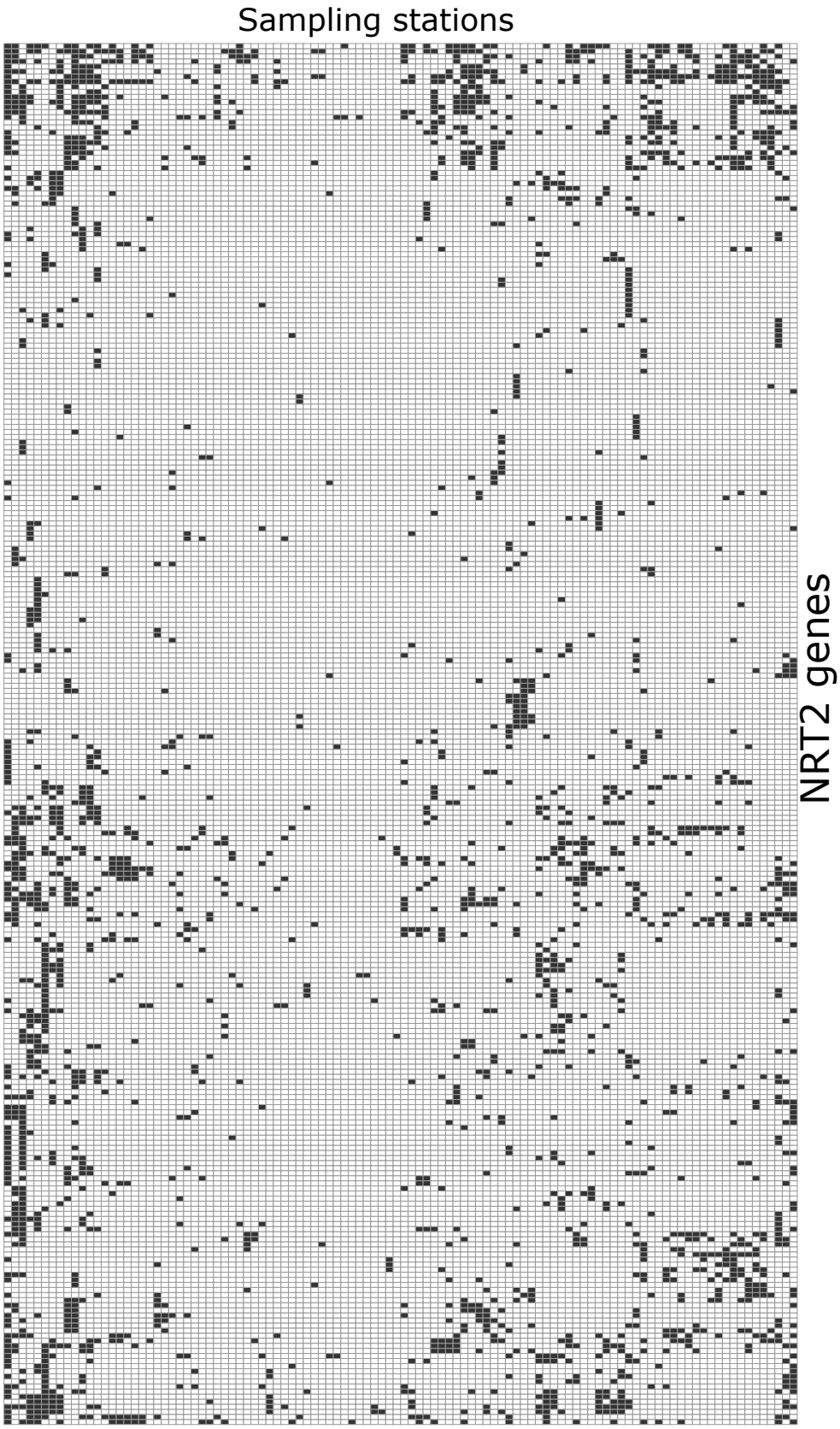

Fig. 3.7: Heatmap of saturation errors occurrences for di-*NRT2*. Grey cells correspond to the stations where the corresponding transporter gene shows this type of error.

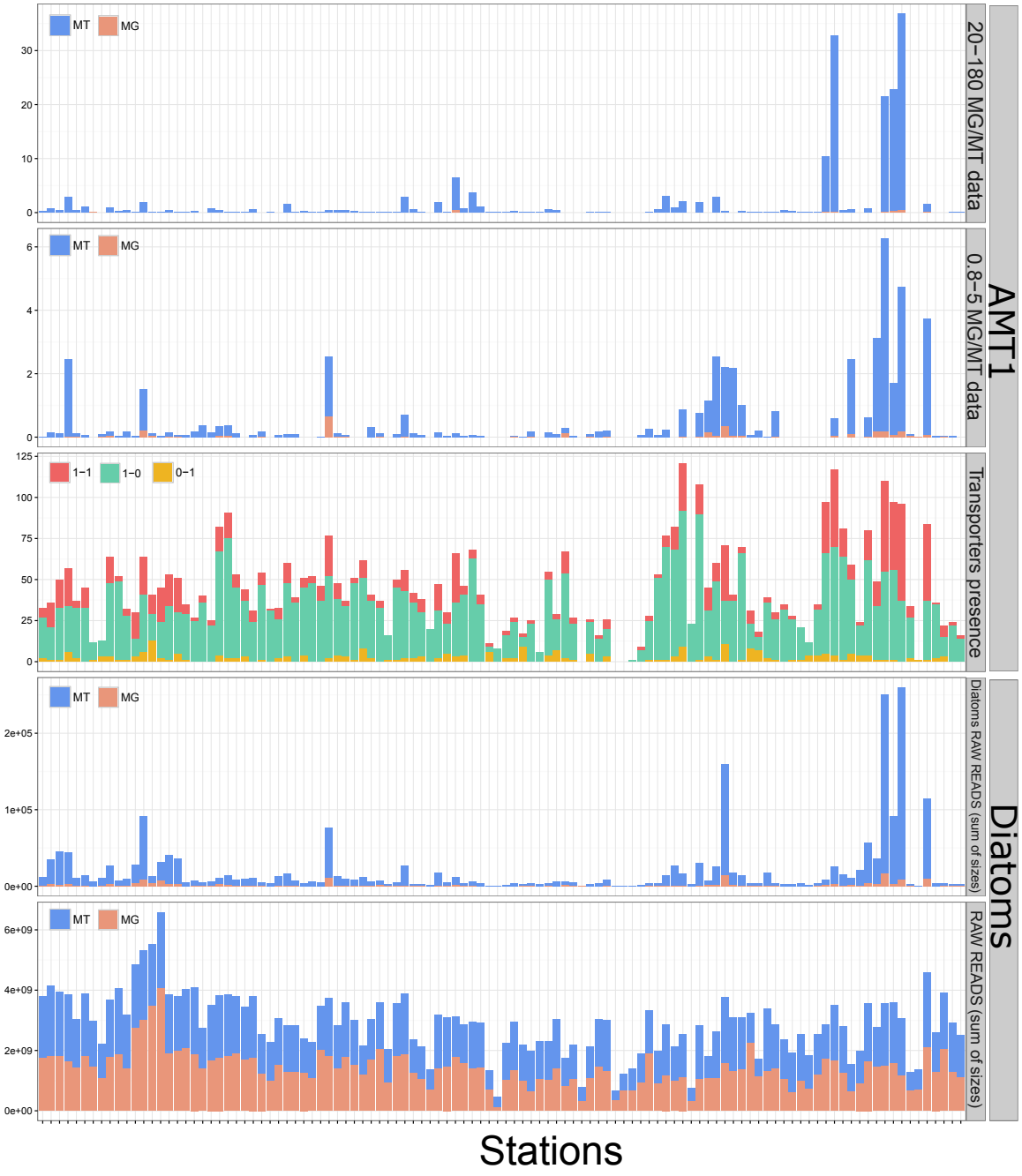

Fig. 3.8: Barplot of the number of *AMT1*-annotated reads sequenced in the metagenomic and metatranscriptomic data (the two top-panels) compared to the number of *AMT1* transporters genes observed in every sample (third panel), coloured according to their presence in both metaG and metaT (1-1, red), present only in the MetaG (0-1, yellow) or present only in the MetaT (1-0, cyan) for two size classes (20-180 μm and 0.8-5 μm). In the fourth panel there is the sum of the number of raw reads of MetaG and MetaT from different size classes assigned to diatoms and in the fifth the whole sum.

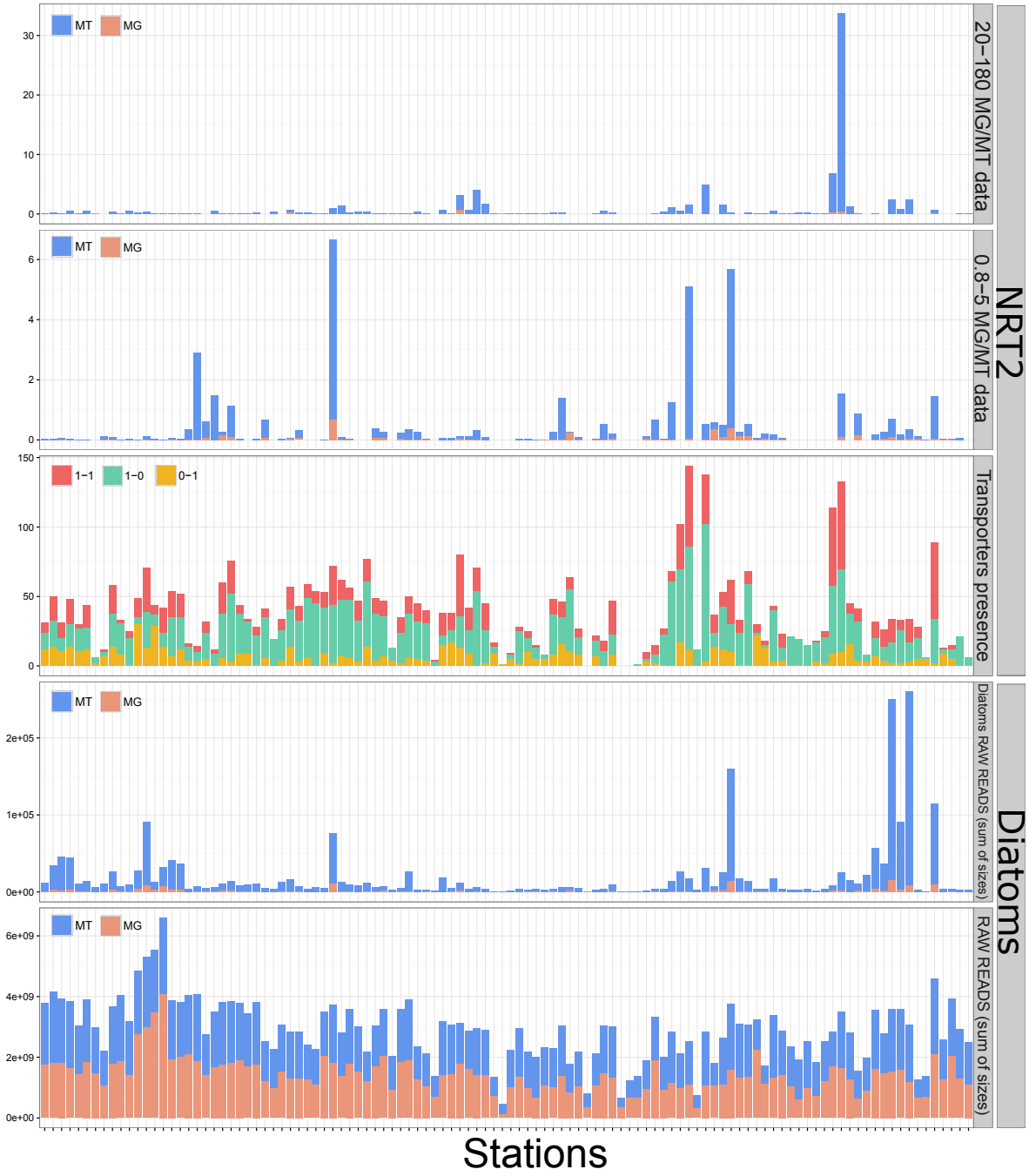

Fig. 3.9: Barplot of the number of *NRT2*-annotated reads sequenced in the metagenomic and metatranscriptomic data (the two top-panels) compared to the number of *NRT2* transporters genes observed in every sample (third panel), coloured according to their presence in both metaG and metaT (1-1, red), present only in the MetaG (0-1, yellow) or present only in the MetaT (1-0, cyan) for two size classes (20-180 μm and 0.8-5 μm). In the fourth panel there is the sum of the number of raw reads of MetaG and MetaT from different size classes assigned to diatoms and in the fifth the whole sum.

As the metagenome was assessed to be error-prone by saturation analysis and the comparison to the corresponding metaT, I preferred to compute the

distribution of every unigene based on the presence-absence observed by the metatranscriptome dataset. The same limits on the metagenome did not allow the normalization of the metatranscriptome over the metagenome dataset. As consequence I normalized the number of transcripts found for every transporter over the total number of reads identified as diatoms in the same metatranscriptomic dataset. This normalization allowed the comparison of expression between different samples, but it is an expression relative to the whole diatom community present in the sample.

Using thus the metatranscriptome information as presence absence I can investigate the distribution of transporter genes. Unigenes are mostly local over the sampling stations: the median percentage of their presence is indeed of 8.5% for di-*AMT1* and of 10.4% for di-*NRT2*, however there is a wide range of unigene ubiquities in both families (Fig. 3.10). Unigenes belonging to di-*AMT1* emerge as ubiquitous genes, reaching ubiquity percentage between 45% and 75%.

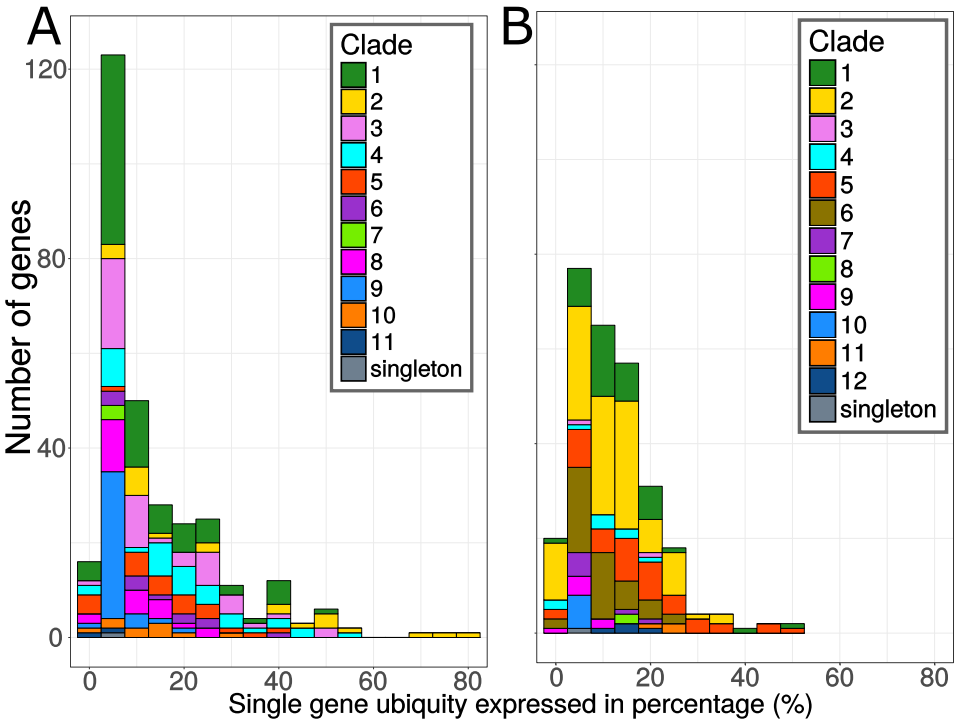

Fig. 3.10: Histogram of unigenes ubiquity. Ubiquity is here expressed as the percentage of stations where each unigene is present over the 106 stations taken into account. Genes are annotated according to their clade assignation.

3.4.6 Clade characterization

The phylogenetic tree clustering resulted in 11 clades for di-*AMT1* and 12 clades for di-*NRT2* with the exclusion of one singleton per gene family not included in any clade. The number of genes composing a clade is extremely variable, from 2 up to 107 genes (Tab 3.3). Ubiquity of unigenes is highly variable, and it is not equally distributed across clades. The median ubiquity of unigenes computed per clade, expressed as percentage of stations covered, ranges from 3.3% of the stations to 23.6%. The coverage of each clade over the *Tara* Oceans stations thus depends not only by the number of genes it is composed by, but also by the ubiquity rate of these same genes (Tab. 3.3).

Ubiquitous clades are found in both families: clade di-*AMT1*-1, di-*AMT1*-2, di-*AMT1*-3, di-*NRT2*-1, di-*NRT2*-2 and di-*NRT2*-5 cover indeed more than 90% of the stations. By contrast, also very local clades were observed, such as di-*AMT1*-7, di-*AMT1*-11 and di-*NRT2*-10, which are present in ~7% of the samples.

Tab. 3.3: Composition and distribution of diatom N transporter clades. For every clade it is here reported the number of genes clustered within the clade and the median ubiquity (in percentage) of the genes belonging to the same clade.

di-*AMT1*	Number of genes	Median ubiquity	di-*NRT2*	Number of genes	Median ubiquity
1	84	6.6	1	42	10.8
2	22	22.6	2	107	10.4
3	50	11.3	3	2	13.7
4	37	17.9	4	9	9.5
5	24	15.6	5	44	17.0
6	12	13.7	6	46	7.5
7	3	4.7	7	7	4.7
8	25	5.7	8	2	13.2
9	37	4.7	9	7	6.6
10	10	12.7	10	7	3.8
11	2	3.3	11	3	23.6
singleton	1	4.7	12	4	15.6
			singleton	1	2.8

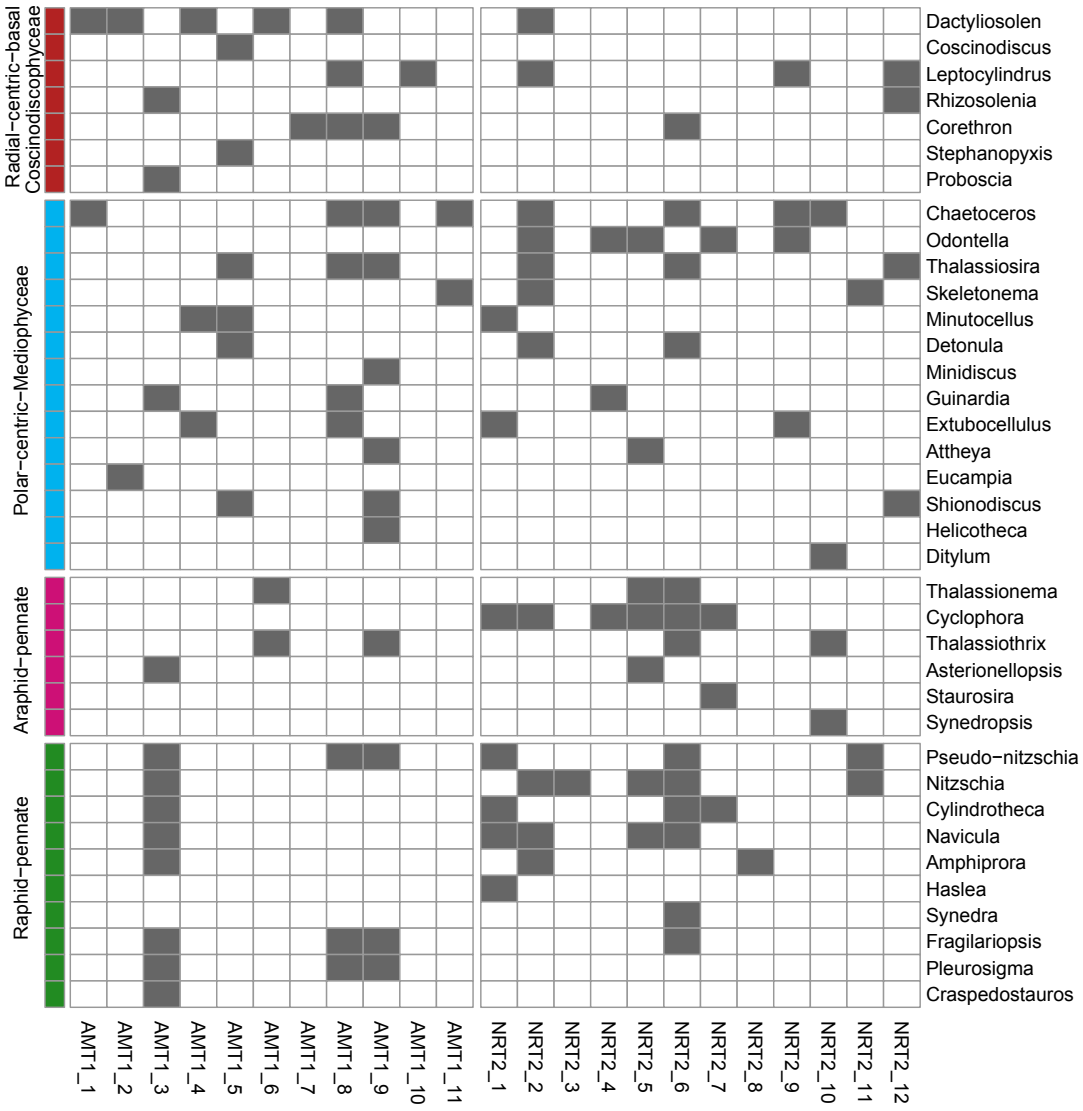

Fig. 3.11: Assignation of diatom genera to the transporters clades based on di-*AMT1* (left panel) and di-*NRT2* (right panel).

Through the taxonomic identification of unigenes I inferred the taxonomical assignation of single clades. Looking at higher taxonomic group annotations (Fig. 3.1 and 3.11) it is clear that clades are widespread across diatoms phylogenies. Some clades are specific of some taxa group while other are shared by all diatoms. Clade di-*AMT1*-3, di-*NRT2*-2 and di-*NRT2*-6 are clades shared by all the diatoms, whose putative function may be interpreted as basal. Clade di-*AMT1*-7 and di-*AMT1*-10 are clades present only in the most basal evolutionary group of diatoms: Radial-centric-basal-Coscinodiscophyceae. Clade di-*AMT1*-11 is the only one belonging only to Polar centric, while clades di-*AMT1*-1, di-*AMT1*-2, di-*AMT1*-4, di-*AMT1*-5, di-*NRT2*-9 and di-*NRT2*-12 are owned by all the centric diatoms. Finally, clades di-*NRT2*-8 and di-*NRT2*-3 are specific of raphid pennate. The fact that clades show some specificity to taxonomic classes is an evidence in favor of the putative functional role they exercise. In the same time, the fact that they are spread across genera also within different groups suggests the absence of taxonomic bias in the definition of these units.

3.4.7 Conclusions

Gene families arise as a consequence of duplication events of an ancestral gene during evolution. These events happen in single individuals and they can be further fixed or lost in the population. The first consequence of duplication is gene redundancy, whereby two or more genes can fulfill the same function (Tautz, 1992). Once duplicated, a given gene can i) degenerate and lose any functional activity, ii) be maintained and maintain its original activity entirely or partially (sub-functionalization), or iii) it can mutate and acquire new functions (neo-functionalization). More specifically, neo-functionalization is the process in which mutations give to one copy of the duplicated gene a new function while sub-functionalization happens when the different copies of a duplicated gene retain a specific sub-function of the original gene each which cumulatively produce the complete original function of the single ancestral

gene (Prince and Pickett, 2002 and references therein). Finally, dosage-balance indicates a survival method for duplicated genes, retaining the same function with a stoichiometric balance in gene expression between the different duplicates (Conant et al., 2014). As results, in any case, within gene families, genes are generally characterized by a slightly different function (which can also be a subfunction), and/or a different localization and/or a differential expression modulation.

It is following this diversity within gene families that I defined, in this chapter two proposals of putative-functional units for diatoms, each based on a N transporter gene family. Contrariwise to previous functional definitions based on a grouping of taxa according to several criteria, the conceptual schema I propose here is completely detached from taxonomy and the concept of species. Diatoms toolkit to handle environmental N is diversified, i.e., it relies on more than one *AMT1* and *NRT2* genes for individual and each of these genes may have a slightly different efficiency in different environmental conditions. The working hypothesis here is thus that each diatom species owns thus a set of genes, each belonging to different clades, and the owning of specific clades allow them to survive in the corresponding environments where it is adapted to (Fig. 3.12). From the evolutionary point of view, the expansion of expression efficiency in a broader range of environmental conditions occurred in turns of gene duplications. Several events of duplication and gene losses have led to the current display of di-*AMT1*- and di-*NRT2*-clades along diatoms tree through evolution. Consequently, the phylogenetic tree of both N transporter gene families taken into account differs completely from diatoms phylogenies. It is thus quite hard to consider these functional units compared to the taxonomy, but this does not fall within the aims of my thesis.

Before investigating the putative functional-role of these clades, which will be addressed in the two further chapters of the thesis (chapters 4 and 5),

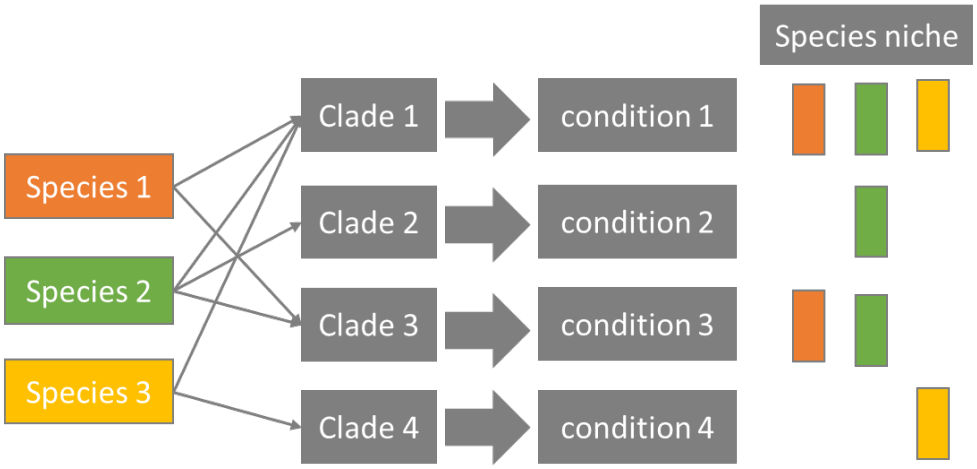

Fig. 3.12: Conceptual scheme of the clades relationship to taxonomic units.

I investigated and described here the characteristics of these clades. I obtain a number of 11 and 12 clades according to di-*AMT1*- or di-*NRT2*- classification, which is three folds as much the number of guilds defined for diatoms within the fresh-water river systems. The higher number is expected as, compared to fresh-water systems, I am now taking into account data from a global-scale sampling, including thus an incredible high number of distinct environmental conditions.

One challenging innovation of my work is the definition of functional units directly over meta-omics datasets, not based on the distribution of genes but rather on a fine evolutionary working-hypothesis that evolutionary close genes share similar ecological functions. The number of N transporter sequences found in the data is far from being comprehensive of all the diatoms di-*NRT2* and di-*AMT1* genes considering the total number of species estimations, and the known number of gene copies species own. However, as the saturation analysis indicates for the metatranscriptomic dataset, it seems that the found N transporter genes actually reach saturation in gene number. This has however to be interpreted as saturation within the sampling strategy designed in the *Tara* Oceans campaign: meaning that areas excluded by this study such as the whole Arctic regions are not taken into account and gave no weight in the saturation analysis to all the species isolated in this area.

Moreover, *Tara* Oceans is a pelagic expedition, I am not taking into account the large proportion of benthic diatoms included in the total estimations of diatoms species number. I can assess thus that within my sampling area I gathered almost the totality of N transporter genes in the metatranscriptome. A further comparison with the total number of species known for the major diatom taxonomy groups and the number of sequences annotated to the same groups depicts how the obtained sequences well reflect the totality of the diatoms phylogeny, displaying the correct relative abundance of species of one group rather than the other.

Even if results are promising for the metatranscriptome dataset, the metagenomic dataset does not exhibit the same promises. Saturation analysis of the metagenome did not reach the plateau, nor for di-*AMT1* genes or for di-*NRT2* genes. Comparing the two datasets a large number of saturation errors resulted, defined as the absence of genes from the metagenome where the corresponding transcript was found in the corresponding metatranscriptome. It is clear hence that metagenome suffers from low coverage and consequent lack of saturation. For this reason I had to consider the metagenome unreliable for the estimation of N transporter gene abundances. This conclusion had two consequences: i) the absence of gene abundances information lead to the use of presence-absence as derived by the metatranscriptome for distribution analysis, and ii) it is not possible to normalize the expression of a gene (derived from the metatranscriptome) over its abundance (derived from the metagenome). Therefore, I chose to normalize the expression of genes over the total number of sequences in the same metatranscriptome datasets annotated as Bacillariophyta.

As I designed the functional units of diatoms over N transporter genes, in the following two chapters I'll investigate the functional diversity distribution and regulation based on this functional definition.

4

Putative functional diversity distribution over the ocean

4.1 Summary and main achievements

- This chapter focuses on the distribution of evolutionary clades as functional units and, consequently, to the estimated functional richness measured over this information;

- Functional richnesses estimated through di-*AMT1* and di-*NRT2* follow very similar patterns, being strongly correlated between the two gene families;

- Compared to the taxonomic richness measured in chapter 2 I observe a different scenario, likely explained by functional redundance of taxa;

- Clades distribution designs global biogeographies dominated by iron, nitrite and nitrate concentration, other than being strongly impacted by latitude.

4.2 Introduction

The limits of taxonomic diversity index in explaining ecosystem functioning increase according to the spatial scale, leading to inconclusive results

at ecosystem levels (Lear et al., 2014). This is due to the fact that we do not have full knowledge on taxa specific functions and dynamics yet. However, where taxonomic diversity fails to correctly understand the processes and the dynamics linking communities to ecosystem functioning, characterizing functional diversity may be the answer (Longhi and Beisner, 2010; Soininen et al., 2016). Functional diversity focuses directly on the function, without having to assign a function to every species, and this bypass allows stronger inferences on the ecosystem state and stability. This change of approach is supported by the massive work done on terrestrial plants, demonstrating that community distribution of traits changes over environmental gradients rather than over geographic distances (e.g., Cornwell and Ackerly, 2009; Swenson and Weiser, 2010), but also by new modeling approaches which proved that community assembly is driven by the series of gene functions available to the community, rather than by the distribution of functions among taxa (Coles et al., 2017). To fully understand which functions allow ecosystem stability and health, in the last decades a massive effort has been made on functional diversity studies, including marine planktonic ones. Among the several branches of ecology, the interest for functional biogeography soon arose to study the geographical distribution of trait diversity. From the distribution of species we are hence going toward the distribution of functions and roles covered by the organisms within and between communities (McGill et al., 2006). Changing perspective can also result in strongly different biogeographical patterns. Ocean prokaryotes composition for example has been observed to be highly variable across regions if observed from a taxonomic point of view, whereas the number of genes classified to each function has been found to be relatively stable (Sunagawa et al., 2015). Focusing on traits distribution rather than species' allows to address fundamental biological questions, such as why a population occupies certain geographic areas and how it would react to environmental changes (Green et al., 2008). Functional biogeography has hence been defined as the study of spatial and temporal distribution of individual functions and of the resulting ecosystems (Reichstein et al., 2014).

Biogeography of phytoplankton communities is mainly driven by adaptation of phytoplankton to the environment, i.e., primarily temperature, nutrient availability, predation and vertical stability (Margalef, 1968; Johnson et al., 2006; Thomas et al., 2012; Tréguer et al., 2018). Nevertheless, very few studies focused on functional biogeography of marine phytoplankton. Temporal distribution of phytoplankton functional units have been investigated by Edwards et al. (2013) exploiting around 7 years of data measured from a station in the western English Channel. They designed five functional traits: three were based on N utilization, one was related to light utilization and one to maximum growth rate. Through this framework they were able to detect seasonal fluctuations in functional terms, proving the valence of traits in predicting community responses under natural conditions, as well as their robustness as interspecific variations indicators. In response to the typical temperate seasonal fluctuations in light and nutrients functional communities showed structural shifts accordingly. Through modeling means Vallina et al. (2014) inferred the functional richness of phytoplankton on a global scale (Fig. 2.4D). They found fast-growing 'opportunist' phytoplankton to dominate the communities at high latitudes, whereas the types able to face limited resource dominate low-latitude regions. The resulting surface functional diversity is lower at polar and subpolar oceans while it is higher in tropical and equatorial regions. Vallina et al. (2014) hence obtained a strong latitudinal gradient, interrupted by hotspots of higher diversity, explained by the authors as peaks of diversity associated to energetic ocean circulation which, because of lateral mixing, works as ecotones where communities of adjacent regions overlap (Barton et al., 2010; Vallina et al., 2014a). The same latitudinal trends have been observed also in taxonomic diversity terms (Chust et al., 2013) and from the satellites information, through specific indices developed over patchiness in ocean color bio-optical anomalies (De Monte et al., 2013). Looking closely to diatoms from a taxonomic point of view Malviya et al. (2016) did not detect a significant latitudinal trends but even if not significant, the patterns

of diversity they observed followed the same trends with a polewards decrease of richness.

Hypothesizing the same drivers of diversity-hotspot formation proposed by Vallina et al. (2014), Lévy et al. (2015) went further in hotspots of functional diversity formation investigating specific fluid structures as well as the phytoplankton diversity associated. Through their numerical model approach they found fronts to be structures especially prone to support high diversity communities. This hypothesis behind the model suggests that fast-growing types are usually the winner of these systems as in these specific topologies nutrients supply are larger and different communities are brought into contact. By contrast within the same system, eddies would have mainly low diversity cores, due to the isolation of communities for times long enough to see predominate the competitive exclusion process (Lévy et al., 2015).

Within this chapter I will take advantage of the previously defined clades as putative functional units of diatoms (chapter 3) to describe the diatoms functional diversity biogeography. Furthermore, through multivariate analysis I will discriminate functional communities and investigate their distribution together with the responsible environmental drivers. This is the first attempt to build a diatom biogeography not based on taxa but on genetic differences of a specific function.

4.3 Material and methods

4.3.1 Data

Transporter abundances and normalization

The abundance of the unigenes selected by the phylogenetic analysis (chapter 3) was extracted by the metatranscriptomic and metagenomic dataset of *Tara* Oceans (Carradec et al., 2018). Occurrences values are computed as the fraction of the number of reads mapped per kb of unigene covered with reads per the total abundance of reads annotated to diatoms in the sample. The sum of occurrences for all the unigenes for a given sample is equal to 1. The dataset of di-*AMT1* and di-*NRT2* occurrences within the metatranscriptome dataset can be found respectively in Supplementary File 9 and 10 while in the metagenomic dataset the data refers to Supplementary File 11 and 12. The total number of reads sequenced per sample is available as well in Supplementary File 13, for the metatranscriptome and in Supplementary File 14 for the metagenome. The data was extracted by the cited public datasets by Dr. Eric Pelletier from Genoscope in Paris.

4.3.2 Data mining

The presence-absence of each clade for both families is defined by the presence in the metatranscriptome database or the metagenome database of *Tara* Oceans. A clades is considered present in a sampling site if the sum of the mRNA abundance of the unigenes belonging to the clade is higher than 0 in at least one of the four size-classes sampled (0.8-5 μm; 5-20 μm, 20-180 μm; 180-2,000 μm). A clade mRNA abundance is computed per each size class and

per sample as the sum of the abundances of unigenes belonging to the same clade.

4.3.3 Transporter richness

As stated previously, in this study the focus is on the transport of inorganic nitrogen into and within the diatom cell, and each phylogenetic clade corresponds to a putative different way for the cell to perform this same function. Transporters richness is expressed for every site as number of evolutionary clades present in the same location. This value has been computed per sample as well as per station, aggregating the presences detected in both sampling depths. The richness of di-*AMT1* and di-*NRT2* has been compared through a Pearson correlation analysis per sampling depth. One-tail t-tests have been performed on the richness values to test if this index in surface is higher than at DCM for both gene families.

4.3.4 Transporter distribution

Zero-adjusted Sørensen dissimilarity coefficient (Eq. 4.1; Clarke et al., 2006) was computed on the clades presence-absence data for the 106 *Tara* Oceans stations through the vegan R package (Oksanen et al., 2017). The coefficient was chosen for its efficiency in dealing with denuded assemblages.

$$D^{Sor-adj} = 100 \cdot \frac{b+c}{2+(2a+b+c)} \tag{4.1}$$

Stations have been clustered applying the Ward's minimum variance method (Murtagh and Legendre, 2014) and aggregated using a cutting levels to form eight clusters of stations (Fig. 4.7). The clustering method choice as well as the optimal cutting level value were supported by the silhouette width

(Rousseeuw, 1987) of the observations using the R package *cluster* (Maechler et al., 2016). To investigate the significant differences between clusters I used permutational multivariate analysis of variance (PERMANOVA; McArdle and Anderson, 2001) applying zero-adjusted Sørensen coefficient similarity matrices built from pairwise comparisons of di-*AMT1* and di-*NRT2* clades presence-absence in every station. The multivariate homogeneity of group dispersions was tested using PERMDISP2 (Anderson, 2006), a multivariate analogue of Levene's test for homogeneity of variances. PERMDISP2 tests if the dispersions (variances) of one or more clusters are different by calculating the distances of clusters members (stations) to the group centroid and subjecting them to a permutation test for homogeneity of multivariate dispersions (PERMUTEST). PERMUTEST performs an ANOVA-like permutation test on the group dispersion and produces pairwise comparisons between groups as a means of post-hoc testing. Both PERMANOVA and PERMDISP2 have been performed in R with the *vegan* package, using 1,000 permutations.

4.3.5 Selection of environmental parameters

Assuming that different functional communities are better fit for specific environmental conditions, I analyzed the distribution of communities in coincidence with a certain suite of environmental conditions. The environmental descriptors of the sites were selected to be key variables a priori related to diatoms N transporter and uncorrelated between them, choosing as threshold a pairwise Pearson correlation lower than = 0.6. The following nine environmental parameters were retained by this selection:

1. Mean chlorophyll α (mg/m^3) as derived from chlorophyll fluorescence sensor, calibrated using HPLC measurements of chlorophyll from concurrent water sampling (Mean_Chloro) (Picheral et al., 2014a);

2. Mean monthly iron concentration (nmol) extracted by the PISCES2 (Aumont et al., 2015) model;

3. Monthly PAR (monthly average ipar) based on satellite data;

4. Monthly mean of surface (5 m depth) NH_4^+ concentration as extracted by World Ocean Atlas 13 (Boyer et al., 2013);

5. $NO_2^- + NO_3^-$ concentration (μmol/l) as extracted *in situ* (Picheral et al., 2014b);

6. Nitrite (NO_2^-) concentration (μmol/l) as extracted *in situ* (Picheral et al., 2014b);

7. Temperature (°C) derived from CTD, SEA-BIRD (Picheral et al., 2014a);

8. Mean nitrocline depth (m) (Picheral et al., 2014a);

9. Sampling depth, categorical variable expressing if the sample was taken at surface (SRF) or at maximum chlorophyll depth (DCM).

In opposition to the other selected variables, chlorophyll α is the only one depicting a biotic information and not a biogeochemical one. Chlorophyll α in this context is utilized as descriptor of the trophic state of the system (Dodds et al., 1998).

4.3.6 Environmental PCA

I performed an environmental Principal Component Analysis (PCA) on all the sampling stations through the R package *FactoMineR* v.1.32 (Lê et al., 2008). A subset of environmental variables was selected specific for each gene

family as the subset of variables better explaining the difference between the communities using the *bioenv* function of the vegan package (Oksanen et al., 2017) (Tab. 4.2). This subset of variables is the one maximizing the Mantel correlation between the dissimilarity matrix of Euclidean distances between stations based on the environmental variables and the dissimilarity matrix of Sørensen distances between the same stations based on the presence absence of the clades. Stations clades-based clusters were mapped on the PCA biplot through a discrete colorimetric scale.

4.4 Results and Discussion

4.4.1 Functional richness

Once defined the functional units, i.e., the clades (see chapter 3), a functional diversity index can be easily measured. Mouchet et al. (2010) identified three main classes of functional diversity measures describing different facets of the same diversity: functional richness, functional evenness and functional divergence. Herein, using N transporter genes as putative functional units I defined functional richness as the number of clades present in each sample (Fig. 4.1).

Patterns of functional diversity over space are visible using both gene families, even if di-*AMT1* richness is overall higher than di-*NRT2*-based one. The Mediterranean Sea and Indian Ocean are characterized by lower diversity as in general it is the case for oligotrophic regions. Hotspots of diversity are located in stations at mid-latitudes such as the Gulf Stream, the Agulhas Current (South Africa), at the Falkland –Malvinas confluence and in the California Current (only di-*NRT2*). These same areas were predicted by modeling means by Barton et al. (2010) to be hotspots of functional diversity. The correspondence of diversity peaks between the trait-based simulations and the

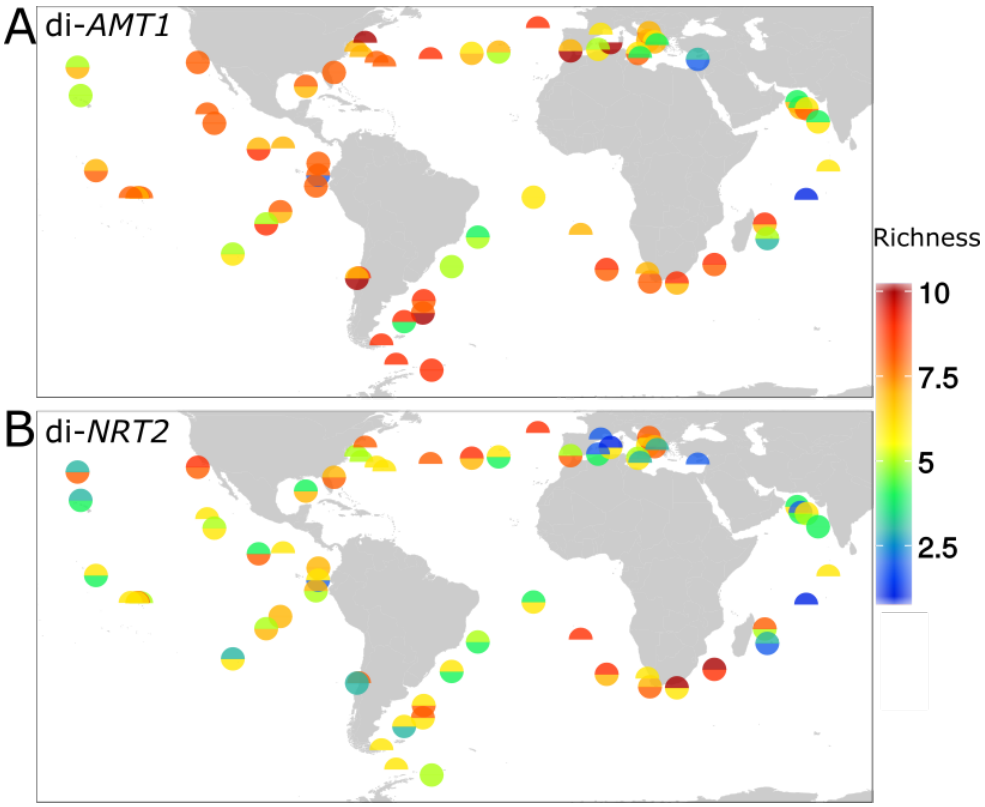

Fig. 4.1: Functional richness expressed as the number of clades present in each station. For both families the maximum number of clades observed in a station is of 10.

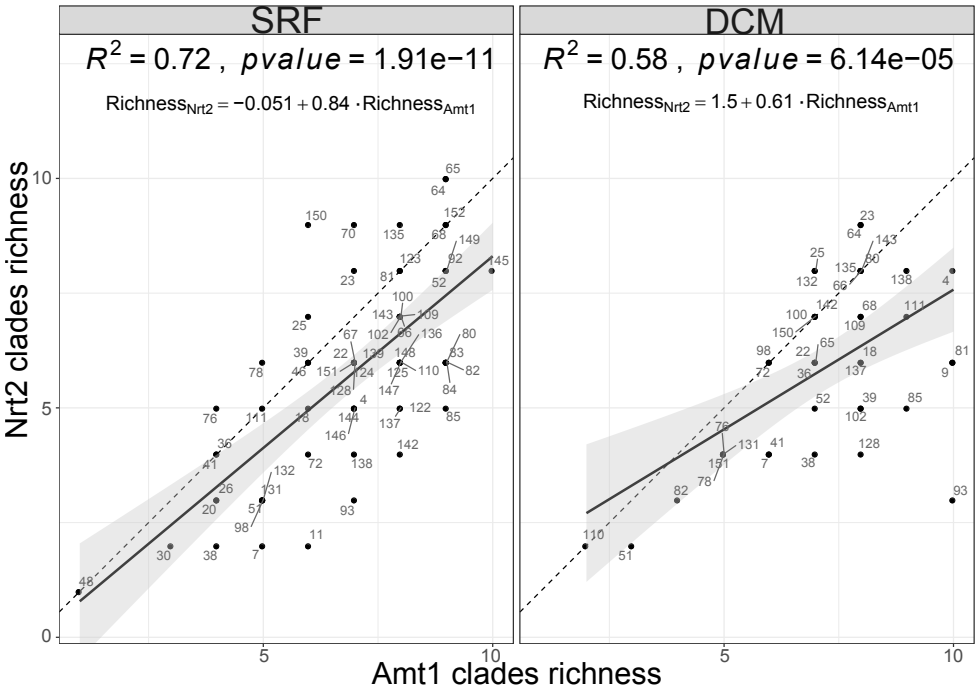

Fig. 4.2: Scatterplot of clade richness as computed on *AMT1* or di-*NRT2* on the two sampling depths and Pearson correlation statistics annotated.

N transporters-based ones supports the putative functional meaning of the functional units here proposed. Interestingly, the largest difference between transporters occurs in nitrate-rich waters, i.e., the iron-limited tropical Pacific and the Southern Ocean, where di-*AMT1* richness shows values comparable with the other hostspots. This feature makes the latter pattern more similar to the more recent modeling exercise by Vallina et al. (2014) and thus suggests that there is a profound functional difference between the two transporters (see also chapter 6). Nevertheless, overall the two gene family based diversities strongly correlates both in surface (rho=0.72, p-value<0.0001) and at DCM (rho=0.58, p-value <0.0001), as depicted by Fig. 4.2.

It is also to note that, where the data is present, at DCM di-*AMT1* richness is relatively higher than di-*NRT2* compared to surface. I will discuss more in details on the differences between SRF and DCM in the following.

Generally, the functional richness herein computed shows several resemblances with the taxonomic richness based on the metabarcoding information

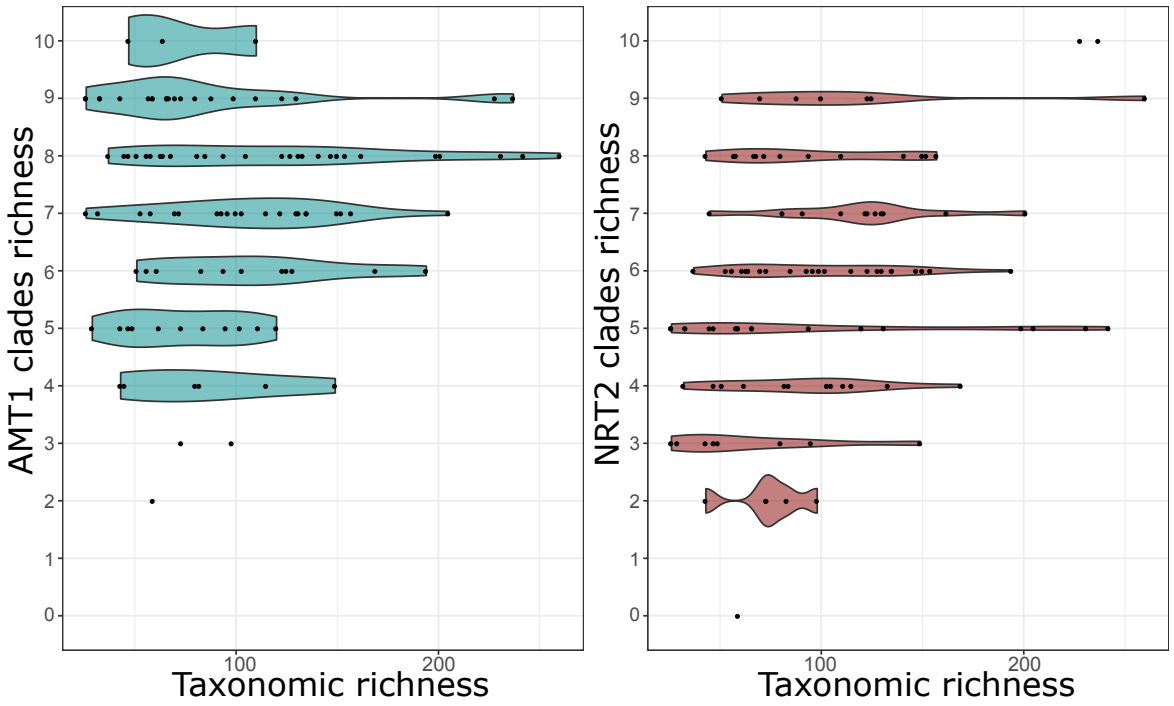

Fig. 4.3: Violin plot of comparison between functional richness as measured by N transporter clades of *AMT1* and *NRT2* gene families and taxonomic richness as obtained by the Swarm-d1 metabarcode of the same stations filtered at 99.65 cumulative abundance threshold, as explained in chapter 2.

(chapter 2). However, directly comparing the three diversity information I did not obtain a linear relationship between the two indices (Fig. 4.3). It is expected for the functional richness to not be linearly related to the taxonomic one. These two information should indeed considerably differ according to the concept of functional redundancy of species. To a low taxonomic richness may correspond very different functional diversity levels indeed. These sites are discriminated by the functional point of view between: communities dominated by few dominant species and communities living in challenging environmental conditions, where only few species survive occupying very different functional roles. While at high taxonomic richness levels we can observe intermediate values of functional richness, suggesting overall a higher number of functional niches displayed but also redundancy of functional roles among species living in the communities.

Tab. 4.1: Multivariate differences between (PERMANOVA) and clusters dispersions within (BETADISPER) based on the presence-absence of the two gene families di-*AMT1* and di-*NRT2*.

	Between		Within	
Cluster family	**F**	**p-value**	**F**	**p-value**
di-*AMT1*	25.84	<0.001 ***	3.39	<0.001 **
di-*NRT2*	13.62	<0.001 ***	1.20	0.3049 (ns)

4.4.2 Biogeography

Clades distribution results both from the number of genes they are composed by, but also by the ubiquity rates of these same unigenes. There are very local clades as well as very ubiquitous ones. Clades such as di-*AMT1*-1, di-*AMT1*-2, di-*AMT1*-3 and di-*NRT2*-2 are present in at least 100 stations over the 106 analyzed while among the local clades there are di-*AMT1*-7, di-*AMT1*-11 and di-*NRT2*-10 which cover just 7% of the sampled stations.

To investigate the distribution of diatom N transporters clades, *Tara* Oceans stations were clustered according to the presence-absence of each clade (Fig. 4.4, Fig. 4.5A) and then mapped in a multivariate environmental PCA (Fig. 4.5). To explore the multivariate differences in clades between different stations clusters, I performed a permutational multivariate analysis of variance (PERMANOVA). This analysis was integrated with a homogeneity of multivariate dispersion test (PERMDISP2; Anderson, 2006) to understand if the distinction between groups was due to a difference in dispersion around group centroids in multidimensional space. The results (Tab. 4.1) proved that, in both families, stations' clusters are significantly different between them, however this difference could be due to different within-group variation in case of di-*AMT1* (Anderson, 2001) as only in di-*NRT2* the null hypothesis of homogeneity between clusters cannot be rejected.

Even if the previous analysis found di-*AMT1*-based clustering to have some weakness in terms of homogeneity within clusters, the clustering method

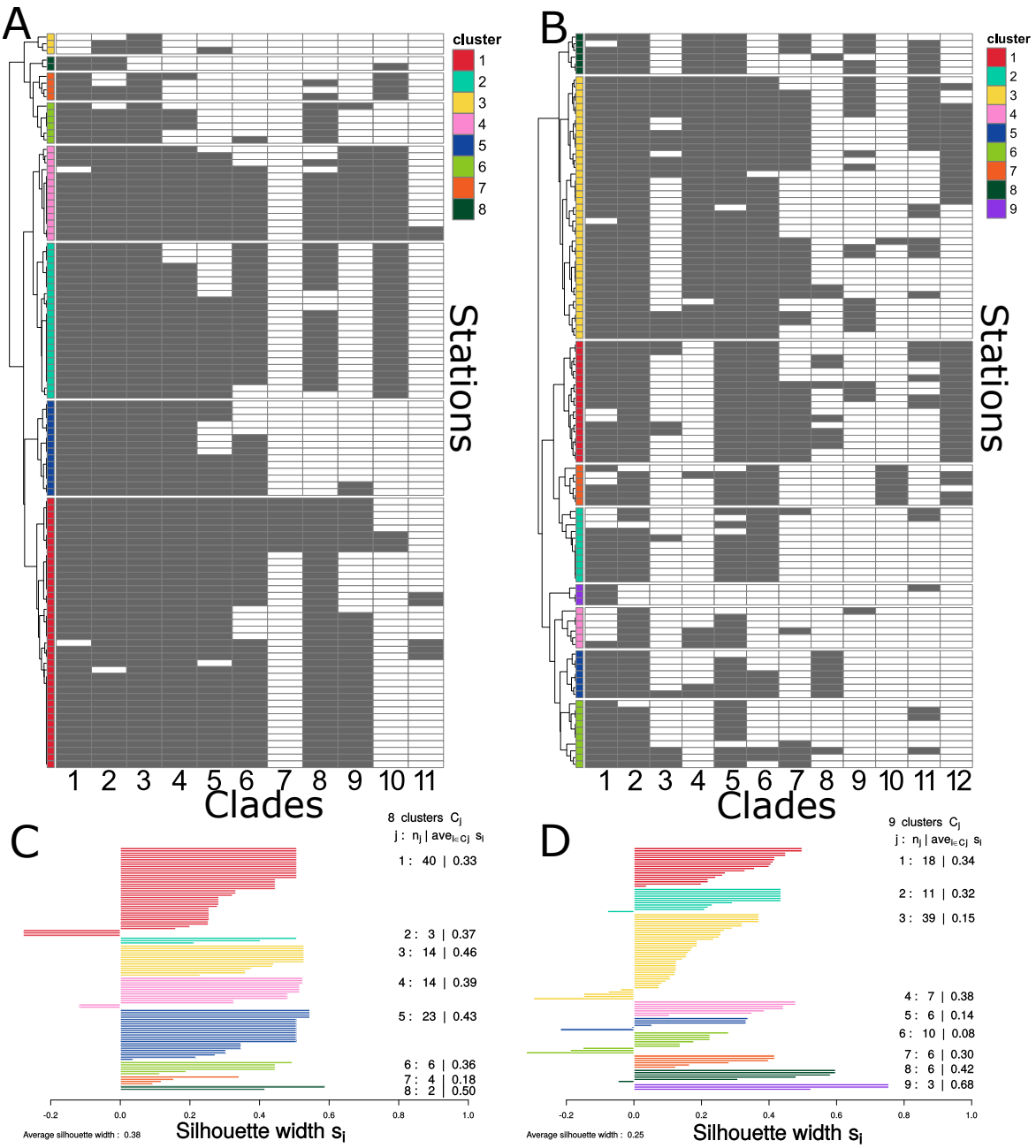

Fig. 4.4: Heatmaps showing di-*AMT1* (A) and di-*NRT2* (B) clades presence-absence and mRNA levels (20-180 μm) in the *Tara* Oceans stations. Stations are clustered by a Ward clustering method based on zero-adjusted Sørensen dissimilarity between the samples based on clades presence absence and annotated in 8 resulting clusters for di-*AMT1* and 9 resulting clusters for di-*NRT2*. The white cells correspond to the stations where the corresponding clade is absent while the gray cells correspond to the stations where the relative clade has been found present in at least one of the size fractions taken into account. In panels C and D are the estimated silhouette values for each established cluster of di-*AMT1* and di-*NRT2*. Values closer to 1 indicate a high degree of similarity of the station within the cluster, positive values close to zero indicate stations which are closer to the other clusters, while negative values indicate stations which may have been misplaced by the clustering. The silhouette defines the clustering as acceptable if all the clusters have elements higher than the average value. The number of sampled stations and the average silhouette value for each cluster are displayed in the right margin.

and cutting level was supported by silhouette analysis (Fig. 4.3 C and D) which showed only few stations within clusters di-*AMT1-red* and di-*AMT1-pink* to have a negative silhouette width, therefore with a weak relationship to the cluster they are located in. Consequently, in the downstream analysis the discussion of the clusters of stations will be addressed taking into account the possible limits of this approach.

Locating the stations within an environmental multivariate space shows di-*AMT1* clustering finding better differentiated structures compared to di-*NRT2* (Fig. 4.5 C and D). The resulting environmental variables better discriminating the clusters are iron and nitrates availability, temperature and light according to the Mantel test (Tab. 4.2). Among these selected parameters both di-*AMT1* and di-*NRT2* clusterings find as main descriptors two different variables having major latitudinal gradients, that is light and temperature, respectively, indicating the presence of major large-scale pattern in the distribution of clades themselves. Only iron availability emerges as a dynamical constraint for both transporters. This environmental parameter derives from modeling and it is thus heavily smoothed at such a global scale. Nevertheless, it is also able to discriminate upwelling areas, areas influenced by iron dust deposition such as the tropical Atlantic Ocean, the Mediterranean Sea and the Indian Ocean. Importantly, the two gene families depend upon two very different forms of nitrates availability: di-*AMT1* the recycled form NO_2^- while di-*NRT2* the $NO_2^- + NO_3^-$ resource.

di-*AMT1* biogeography

A similar clustering, as the one observed for di-*AMT1*, was observed by Malviya et al. (2016) on a subset of the same *Tara* Oceans dataset, whose biogeography was based on the percentage of ribotypes shared between stations. In both clusterings (di-*AMT1-red*) Antarctic stations are close to Southern Africa

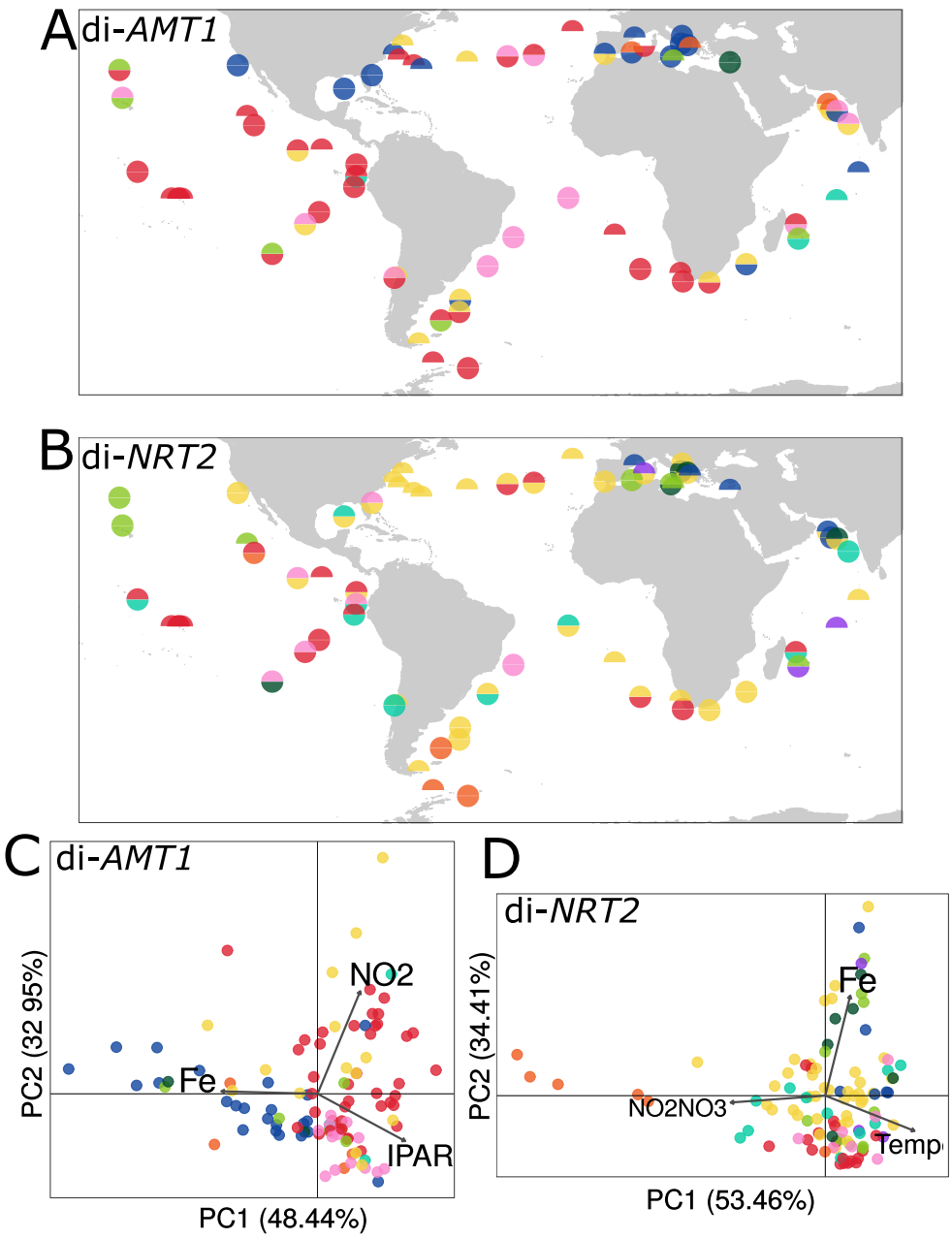

Fig. 4.5: Geographical clusters based on di-*AMT1* (A) and di-*NRT2* (B) clades presence/absence. The top portion of each circle represents samples collected at the surface and the bottom portion represents the DCM (stations missing metatranscriptome data for one of the two depths are drawn as half circles). The biplots of the environmental PCA of di-*AMT1* (C) and di-*NRT2* (D) with a density contour of the clusters previously defined. Each point corresponds to a sampled station while the arrows correspond to the descriptors of the PCA space. Eight clusters result from di-*AMT1* data whereas nine clusters from di-*NRT2*. Clusters are identified by up to nine colors per family: *yellow*, *cyan*, *pink*, *blue*, *red*, *green*, *orange* and *dark green* and *violet*. Clusters are defined in Fig. 4.4 (see methods).

stations (stations 65 and 67), probably as consequence of dispersal between the Antarctic Circumpolar Current and the Agulhas current, but also due to the similar environmental conditions of the two regions. These are characterized by high nitrate availability and coherently this cluster is found also in the nitrate-rich tropical Pacific. Other similarities between the di-*AMT1* and Malviya et al. (Malviya et al., 2016) clusterings include a coherence within Mediterranean stations (di-*AMT1-darkgreen*), and also the close relationship between Mediterranean and Indian Ocean stations (di-*AMT1-orange*). The mid-latitude Pacific stations are clustered together, close to the Antarctic stations for Malviya, and I obtain the same patterns, with all the same stations clustered together in di-*AMT1-red* a cluster dominated by low iron- and high NO_2^- availability. Another relevant result from di-*AMT1* clustering is the closeness of Mediterranean Sea samples with the western Atlantic stations, characterized by high iron availability (di-*AMT1-blue*), and the fine clustering of nutrient-limited tropical stations (di-*AMT1-pink*) and oligotrophic areas (di-*AMT1-green*).

Tab. 4.2: Bioenv output for the selection of the environmental parameters for the Principal Component analysis on presence-absence data.

Environmental parameters subset	**Correlation**
di-*AMT1*	
Fe	0.1833
Fe, Monthly_ipar	0.1934
Fe, Monthly_ipar, NO_2^-	**0.2326**
Mean_Chloro, Fe, Monthly_ipar, NO_2^-	0.2251
Fe, Monthly_ipar, NO_2^-+NO_3^-, NO_2^-, Nitrocline	0.2005
Fe, Monthly_ipar, NO_2^-+NO_3^-, NO_2^-, Temperature, Nitrocline	0.1797
Mean_Chloro, Fe, Monthly_ipar, NO_2^-+NO_3^-, NO_2^-, Temperature, Nitrocline	0.1547
Mean_Chloro, Fe, Monthly_ipar, NH_4^+, NO_2^-+NO_3^-, NO_2^-, Temperature, Nitrocline	0.1184
di-*NRT2*	
Fe	0.1532
Fe, NO_2^-+NO_3^-	0.2100
Fe, NO_2^-+NO_3^-, Temperature	**0.2103**
Fe, NO_2^-+NO_3^-, NO_2^-, Temperature	0.2044
Fe, Monthly_ipar, NO_2^-+NO_3^-, NO_2^-, Temperature	0.1903
Mean_Chloro, Fe, Monthly_ipar, NO_2^-+NO_3^-, NO_2^-, Temperature	0.1756
Mean_Chloro, Fe, Monthly_ipar, NH_4^+, NO_2^-+NO_3^-, NO_2^-, Temperature	0.1445
Mean_Chloro, Fe, Monthly_ipar, NH_4^+, NO_2^-+NO_3^-, NO_2^-, Temperature, Nitrocline	0.1106

By contrast, di-*NRT2* clustering sees rather high divergences with the biogeography designed by di-*AMT1*-data. di-*NRT2*-clustering depicts a wide cluster of stations, di-*NRT2-yellow*, widespread through all the Atlantic and Mediterranean stations, with a remarkable latitudinal symmetry. Smaller clusters of stations are confined to other specific regions: di-*NRT2-orange* isolates the Antarctic stations, di-*NRT2-green* depicts all the oligotrophic samples, di-*NRT2-blue* defines the typical Mediterranean composition, shared by the north Indian Ocean samples and finally to di-*NRT2-red* belongs the majority of the mid-latitude Pacific stations and few South-East Atlantic stations, characterized by low N and iron availabilities and medium-high temperatures, similarly to what observed for di-*AMT1*.

4.5 Conclusions

The biogeography of microorganisms is an issue debated since the last century. As the related data are hard to access and collect, several theories have been postulated over the possible drivers, limits or even lack of limits of micro-organisms distribution, without having the appropriate means to strongly support any position (see chapter 1.2). Indeed, even if the first theories were proposed in the 20s, the first evidences were published only in the last decades, leaving still the place for several uncertainties. Moreover, if we focus on functional biogeography rather than on the taxonomic one, we still are at the dawn of this very promising field and much work remains to be done. Indeed, for phytoplankton we mainly achieved modeling data up to now to describe its global biogeography (Barton et al., 2010; Vallina et al., 2014a). *Tara* Oceans expeditions proposed a promising set of data to meet the needs of this ecological question. Malviya et al. (2016) already investigated the biogeography of diatoms from a taxonomic point of view exploiting this same

dataset. They found diatoms genera to follow three main patterns at global scale: i) they could show lower diversity in the tropics (e.g., *Fragilariopsis*), or ii) they had lower diversity at higher latitudes (e.g., *Guinardia*) or iii) they rather presented an overall uniform diversity (e.g., *Thalassiosira*).

In this chapter I proposed a new functional biogeography of the same phytoplanktonic group. Exploiting the putative functional units designed over the evolutionary clades of N transporter genes (chapter 3), I hereby investigated the distribution of the consequent functional richness and of the corresponding functional assemblages.

Compared to the taxonomic richness I estimated in chapter 2, the richness measured in this chapter showed marked differences. As these two points of view of diversity conceptually differ it is expected to find discrepancies between the two, in particular due to taxa plasticity and functional redundancy. Indeed, the taxonomic richness hotspots were characterized by intermediate functional richness, and this may be explained by functional redundancy. For what concerns functional hotspots, they do not strongly correspond to high taxonomic richness stations. Instead, these hotspots may correspond to more stable ecosystems where a higher specialization of taxa lead to a high number of functional niches.

The putative functional richness I designed well reflects what has been modeled by Barton et al. (2010) and by Vallina et al. (2014), detecting all the modeled main hotspots of diversity. This is the first evidence in favor of the functional information carried by N transporter genes. These genes families provided a quite similar information. The uptake solutions adopted by diatoms may be thus be considered as a good descriptor of the resource utilization trait. This trait has been widely used as main descriptor of phytoplankton functional diversity (Edwards et al., 2013a; Edwards et al., 2013b) and it is routinely considered in modeling exercises (e.g., Barton et al., 2010). It is a

trait explicative of the response of an organisms to environmental conditions and for this reason it can be classified as response traits (Suding et al., 2008), but at the same time it also indicates the direct effects of microorganisms over biogeochemical processes and can be classified as effect traits (Litchman et al., 2015b). Being both a response and an effect trait, nutrient utilization traits have a very comprehensive explicative power, providing with a single information clues about biogeochemical processes (Lavorel and Garnier, 2002), microbial community structures and the environment (Litchman et al., 2015b). Given the relevance of this trait and the difficulty to measure it *in situ*, Litchman et al. (2015) encouraged the research of genomic signatures of resource utilization traits, suggesting as example the same di-*NRT2* genes I selected. I cannot assess with my data the real functionality of the clades, as laboratory experimentations are compulsory to provide a full functional characterization. Given the data I had access to, the only way to validate this approach is to assess the reasonability of the resulting patterns and indeed the biogeography of functional richness follows what it was to be expected according to numerical modeling predictions. The comparison between omic and modeling approaches can lead to important findings, a further step in this direction is presented in chapter 6.

Looking at the functional assemblages assessed by the clustering of stations over the clades presence I obtain further information on the biogeography of this function. The geographic distribution of these clusters partially reflects the nitrate availability, but it is clear, by the nitrate availability maps (Fig. 4.6), that N alone cannot justify the distribution of stations clustering. The PCA performed on these clusters showed clusters to occupy different biogeochemical areas in terms of nutrient availability (iron and N) as well as major latitudinal parameters such as light and temperature.

The fact that at least one parameter with strong latitudinal gradients is retained as significant for each gene-family based clustering highlights the

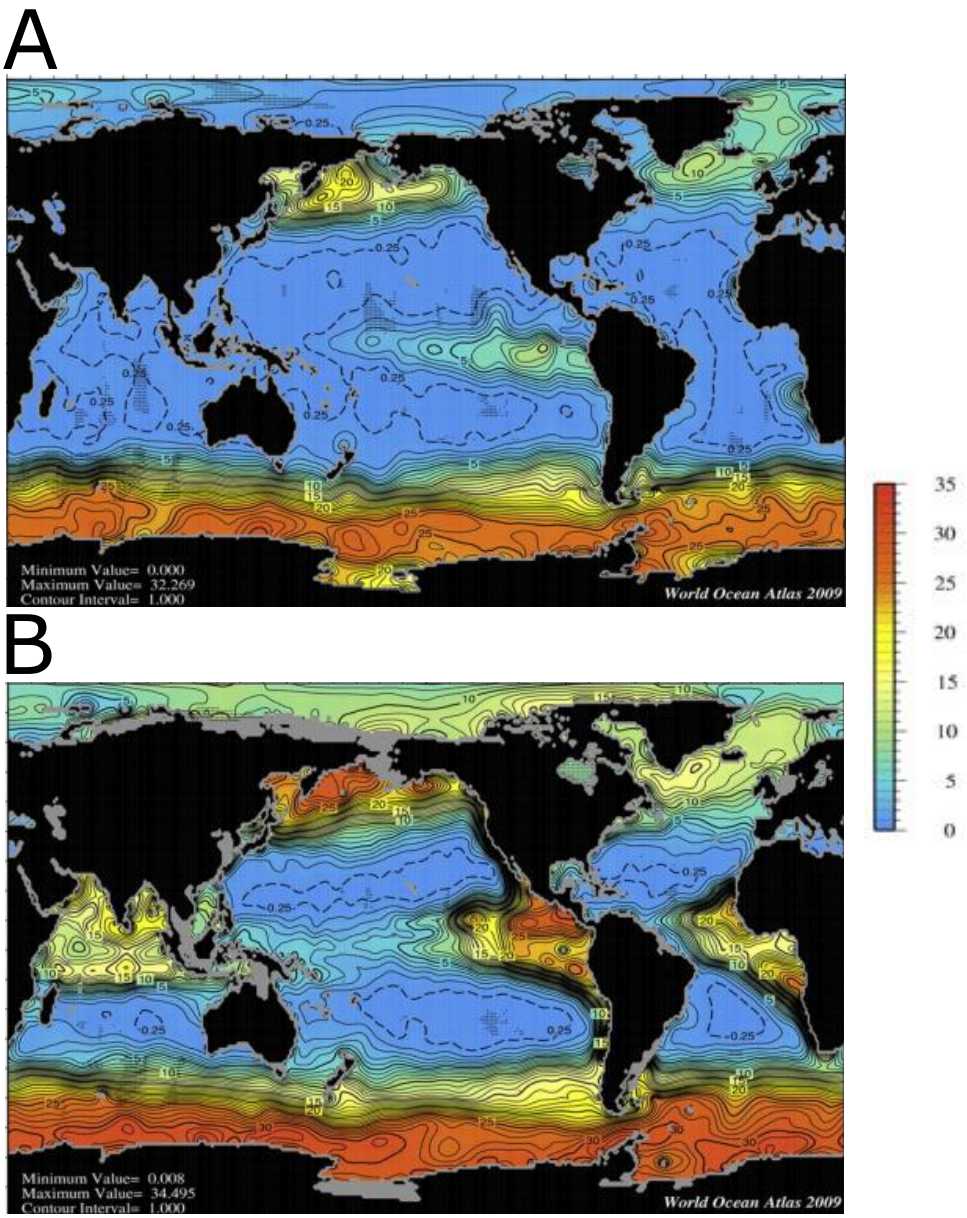

Fig. 4.6: Average spatial concentration (μm/L) of nitrate in the ocean. A: NO_3^- at surface; B: NO_3^- at 100 m.

impact of this particular gradient on functional communities. As already seen in other contributions (Barton et al., 2010; Chust et al., 2013; De Monte et al., 2013; Sunagawa et al., 2015; Malviya et al., 2016) latitudinal gradients are fundamental in structuring phytoplanktonic structures both from a taxonomic and a functional point of view. With my results I confirm the presence of this pattern also for diatoms in functional terms, even if it is not the only and main driver of diversity distribution. Indeed, while taxonomic diversity is strongly influenced by geography, functionally regions are not discriminated according to their geographic distance but only by their environmental conditions. In agreement with this statement, the environmental variables found to be the major describers of functional communities discriminations are nutrients such as N sources and iron availability.

The presence of N sources as driver of communities distribution is expected, being the functional communities based on genes involved in the uptake of the same nutrient. Indeed, diatoms are expected to adopt specific clades of transporters to face different nitrates and ammonium availabilities and the fact that parameters related to two specific, different N availabilities have been selected are proof of the quality of clades as representatives of the resource utilization trait. The second nutrient arising as significant in the communities discrimination is iron availability, as modeled by the global ocean model PISCES2 (Fig. 4.7).

The presence of a tight relationship between iron availability and N uptake is already known. Indeed, Cohen et al. (2017) found both *AMT* and *NRT* transporters genes to be strongly modulated over the stress induced by iron limited availability. The linkage between iron and N metabolisms makes the N transporter genes not only good describers of N availability but also of iron availability.

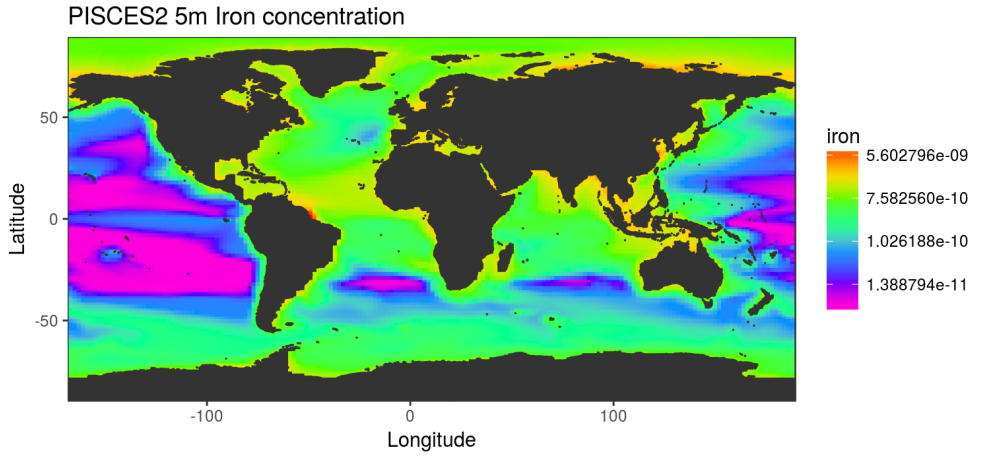

Fig. 4.7: Average spatial concentration of iron in the ocean as modeled by PISCES2 at surface.

To conclude, while in diversity terms the two transporters gave very similar patterns, focusing on functional communities unveiled slightly different patterns. Di-*AMT1*-based communities seem indeed to better discriminate different biogeochemical areas, exhibiting a more defined distribution over the environmental multivariate space, while the di-*NRT2*-based communities are rather overlapped in their environmental characterization. This was also reflected in the different richness patterns, where the first showed a much higher diversity in iron-limited regions.

5

Environmental-based modulation of putative functional units

5.1 Summary and main achievements

- In this chapter I determinated the functional role of different transporters evolutionary clades. This was estimated indirectly exploiting their modulation in relation to the environmental context where they are preferentially found;

- The di-*NRT2* gene family, taken as a whole, sees its mRNA level to be clearly modulated by N availability, having higher mRNA levels the lower is the concentration of their substrate. The answer is less clear for di-*AMT1*, which follows higher geographical patterns, being more linked to the abundance of diatoms themselves in the community;

- There is a clear regionalization of the relative mRNA abundance of single transporter clades, delineating in particular specific isolated areas as the Mediterranean Sea or the Southern Sea;

- The majority of clades is strictly linked to one size class rather than the others, highlighting the specific evolutionary solutions adapted by differentially sized diatoms to face N uptake. This strenghtens the proposed definition of functional units over metatranscriptomic data, as

the designed units are able to capture also the different sizes, being size a fundamental functional trait of phytoplankton;

- There is a strong vertical gradient in the use of di-*AMT1* or di-*NRT2*, peculiarly reflecting the different N sources available at surface rather than at DCM. di-*AMT1* clades in particular seem to be linked to the transcriptomic activity of N metabolism-related prokaryotes modules, suggestive of a similar regulation or rather to the use by diatoms of public goods released by the prokaryotic activity;

- I found with multivariate and machine learning means that iron availability, nitrogen sources concentration and temperature are the main cues defining the functional units distribution;

- Boosted regression trees modeling proved to be a great tool in defining the optimal condition for the use of each clade. I was able through this modeling to indirectly delineate the functional role of different clades;

- Concerning the climate change context, the same modeling tool allowed to predict the answer of clades to several temperature increase scenarios, leading to a dramatic distribution decrease of specific clades.

5.2 Introduction

As a premise to the methodological approach here developed lies the hypothesis that clades own different putative functional roles. As no laboratory experiments could be performed on this data to investigate this hypothesis, a different functional differentiation is herein proposed based on the investigation of differential gene expression regulation. Indeed, to face the rapid changes of environmental conditions, which are typical of the planktonic

lifestyle, diatoms have to rigorously regulate N transporter genes (Rogato et al., 2015). Within this context, N transporter clades modulation might be considered informative of different adaptations to the use of N. Consequently, this information can be exploited as clades functional describer.

Literature already described how diatom ammonium and nitrate transporters genes can be modulated over several parameters (Tab. 5.1 and 5.2).

Through PCR methods NRT2 from diatoms species *Chetoceros affinis*, *Chaetoceros muelleri*, *Skeletonema costatum*, *Cylindrotheca fusiformis* and *Thalassiosira weissflagii* were all found to be overexpressed in nitrogen starvation conditions and underexpressed in NH_4^+ high availability (Song and Ward, 2007; Liu et al., 2013; Kang et al., 2015). Similar regulation results were obtained also for *AMT1* genes through gene cloning in *C. fusiformis* (Hildebrand, 2005). Elaborating deeper methods as transcriptomics and analysing transcripts expression of each gene lead to differential expression answers at gene levels. For what concerns NRT2, *Thalassiosira pseudonana*, *Phaeodactylum tricornutum* and *Fragilariopsis cylindrus* presented among the inducible genes only overexpressed *NRT2* genes in N starvation conditions (Mock et al., 2008; Bhadury et al., 2011; Bender et al., 2014; Levitan et al., 2015; Alipanah et al., 2015) while in *Pseudonitzschia multiseries* there are genes both over-expressed and under-expressed in the same stressting condition (Bender et al., 2014). For *AMT1* results are more complex in N starvation: the previous observations of overexpression were corroborated in *F. cylindrus* and *P. multiseries* (Bender et al., 2014) while in *T. pseudonana* inducible genes were underexpressed (Bender et al., 2014; Ashworth et al., 2013) and in *P. tricornutum* both over-expressed and under-expressed genes were found.

Other than nitrogen availability, other internal and external cues have been suggested to play a role in nitrate and ammonium uptake transcriptional regulation. Light (Ashworth et al., 2013; Bender et al., 2014), temperature

(Lomas and Glibert, 1999) and turbulence (Amato et al., 2017; Dell'aquila et al., 2017) were found to influence uptake expression as well as main nutrients availability as phosphate (Liu et al., 2013), silicate (Sapriel et al., 2009) and iron (Cohen et al., 2017). Other than external cues also internal processes like the cell cycle has been suggested to have a role in N uptake regulation (Hildebrand and Dahlin, 2000) like the internal nutritional status of the cell (Allen et al., 2011).

Tab. 5.1: High affinity Ammonium transporters modulation information from literature. If the literature underwent transcriptomic analysis the modulation information refers to single proteins, while if the analysis were not done at the single-protein level the protein ID is not indicated.

Diatom	Data	Protein	Expression data
Thalassiosira pseudonana	Transcriptomic	36263	-
	Transcriptomic	40537	Regulated by light (Ashworth et al., 2013)
	Transcriptomic	14096	Downregulated in N starvation (Ashworth et al., 2013; Bender et al., 2014) and downregulated in Si starvation (Smith et al., 2016)
	Transcriptomic	13996	Downregulated in N starvation (Ashworth et al., 2013; Bender et al., 2014) and downregulated in Si starvation (Smith et al., 2016)
	Transcriptomic	268793	Downregulated in N starvation (Ashworth et al., 2013; Bender et al., 2014) and upregulated in Si starvation (Smith et al., 2016)
	Transcriptomic	268226	Upregulated in Si starvation (Smith et al., 2016)
	Transcriptomic	258067	-
Thalassiosira rotula	MetaT mapped to KO gene families	-	Downregulated in P starvation (Alexander et al., 2015)
Phaeodactylum tricornutum	Transcriptomic	27877	Upregulated in N starvation (Valenzuela et al., 2012; Levitan et al., 2015) and Si availability (Sapriel et al., 2009) but downregulated in P starvation (Alipanah et al., 2018)

	Transcriptomic	1862	Upregulated in N starvation (Levitan et al., 2015)
	Transcriptomic	1813	-
	Transcriptomic	10881	-
	Transcriptomic	13418	Upregulated in N starvation (Levitan et al., 2015)
	Transcriptomic	54981	Downregulated in N starvation (Alipanah et al., 2015) and in P starvation (Alipanah et al., 2018)
	Transcriptomic	11128	Downregulated in N starvation (Levitan et al., 2015)
	Transcriptomic	51516	Upregulated in N starvation (Alipanah et al., 2015) and downregulated in P starvation (Alipanah et al., 2018)
Fragilariopsis cylindrus	Transcriptomic	212054	-
	Transcriptomic	275907	Upregulated in N starvation (Bender et al., 2014)
	Transcriptomic	209552	-
	Transcriptomic	234670	-
	Transcriptomic	225282	Upregulated in N starvation(Bender et al., 2014)
	Transcriptomic	230306	-
	Transcriptomic	195049	Upregulated in N starvation Bender et al., 2014
	Transcriptomic	257791	-
Pseudo-nitzschia multiseries	Transcriptomic	41386	-
	Transcriptomic	240985	Upregulated in N starvation (Bender et al., 2014)
	Transcriptomic	258485	Upregulated in N starvation (Bender et al., 2014)
	Transcriptomic	256647	-
	Transcriptomic	286729	-

Cylindrotheca fusiformis	PCR amplification	-	Upregulated in N starvation and downregulated in NH_4^+ supplemented (Hildebrand, 2005)
Chaetoceros decipiens	Transcriptomic mapped to GO	-	Downregulated in turbulence conditions (Amato et al., 2017)
Skeletonema spp.	MetaT mapped to KO gene families	-	Downregulated in N starvation (Alexander et al., 2015)
Skeletonema marinoi	Transcriptomic mapped to GO	-	Upregulated as indirect effect of grazing (Amato et al., 2018)

Tab. 5.2: High affinity Nitrate transporters modulation information from literature. If the literature underwent transcriptomic analysis the modulation information refers to single proteins, while if the analysis were not done at the single-protein level the protein ID is not indicated.

Diatom	**Data**	**Protein**	**Expression data**
Thalassiosira pseudonana	Transcriptomic	27414	Upregulated in N starvation (Mock et al., 2008; Bhadury et al., 2011; Bender et al., 2014), downregulated in P starvation (Dyhrman et al., 2012) and downregulated in Si starvation (Smith et al., 2016)
	Transcriptomic	269274	Upregulated in N starvation (Mock et al., 2008; Bender et al., 2012; Bender et al., 2014), downregulated in P starvation (Dyhrman et al., 2012) and downregulated in Si starvation (Smith et al., 2016)
	Transcriptomic	39592	Regulated by light (Ashworth et al., 2013) and downregulated in Si starvation (Smith et al., 2016)
Thalassiosira rotula	MetaT mapped to KO gene families	-	Downregulated in N or P starvation (Alexander et al., 2015)

Thalassiosira weissflagii	PCR amplification	-	Upregulated in N starvation and downregulated in NH_4^+ supplemented (Song and Ward, 2007)
Phaeodactylum tricornutum	Transcriptomic	54101	Upregulated in N starvation (Valenzuela et al., 2012; Alipanah et al., 2015; Levitan et al., 2015) and downregulated in P starvation (Alipanah et al., 2018)
	Transcriptomic	26029	Upregulated in N starvation (Levitan et al., 2015) and downregulated in urea or NH_4^+ supplemented (Allen et al., 2011) and in P starvation (Alipanah et al., 2018)
	Transcriptomic	54560	Upregulated in N starvation (Alipanah et al., 2015; Levitan et al., 2015) and downregulated in P starvation (Alipanah et al., 2018)
	Transcriptomic	2032	Downregulated in P starvation (Alipanah et al., 2018)
	Transcriptomic	2171	Downregulated in P starvation (Alipanah et al., 2018)
	Transcriptomic	40691	Downregulated in P starvation (Alipanah et al., 2018)
Fragilariopsis cylindrus	Transcriptomic	174569	-
	Transcriptomic	229932	Upregulated in N starvation (Bender et al., 2014)
	Transcriptomic	263088	Upregulated in N starvation (Bender et al., 2014)
	Transcriptomic	229065	Upregulated in N starvation (Bender et al., 2014)
	Transcriptomic	251355	-
	Transcriptomic	229096	Upregulated in N starvation (Bender et al., 2014)
	Transcriptomic	251307	-
	Transcriptomic	249557	-
	Transcriptomic	251306	-

Pseudo-nitzschia multiseries	Transcriptomic	261779	Downregulated in N starvation (Bender et al., 2014)
	Transcriptomic	293087	Upregulated in N starvation (Bender et al., 2014)
	Transcriptomic	325874	Upregulated in N starvation (Bender et al., 2014)
Skeletonema costatum	PCR amplification	-	Upregulated in N starvation and downregulated in NH_4^+ supplemented (Song and Ward, 2007; Kang et al., 2015)
Skeletonema marinoi	Transcriptomic mapped to GO	-	Upregulated as indirect effect of grazing (Amato et al., 2018)
Skeletonema spp.	MetaT mapped to KO gene families	-	Upregulated in N supplemented (Alexander et al., 2015)
Chaetoceros muelleri	PCR amplification	-	Upregulated in N starvation and downregulated in NH_4^+ supplemented (Song and Ward, 2007)
Chaetoceros affinis	PCR amplification	-	Upregulated in N starvation and downregulated in NH_4^+ supplemented (Kang et al., 2015)
Chaetoceros decipiens	Transcriptomic mapped to GO	-	Upregulated in turbulence conditions (Amato et al., 2017)

Moreover, the transcription level of a gene is a necessary but not sufficient condition to assess the downstream increase of the corresponding protein in the cell: other levels of control are indeed at stake. Even if up to now regulation of N transporter genes has been studied mostly at the transcriptional level, an additional level of control such as post-transcriptional regulations has been suggested. This has been hypothesized through indirect evidence for *AMT1* by Rawat et al. (1999) and Kumar et al. (2003) from observation in *Arabidopsis* and *Oryza sativa* respectively. Post-trascriptional regulation of

NRT2 genes has been firstly suggested by Fraisier et al. (2000) in *Nicotiana plumbaginifolia* and then directly proved on *NRT2* by Laugier et al. (2012) studies on *A. thaliana*. This latter study stated the predominant role of this second regulatory level over the transcriptional one, even if they observed also the contribution of transcriptional regulation under specific conditions. The regulatory system of N transporters in diatoms is thus expected to be complex and to work at different levels.

From the phylogenies of di-*AMT1* and di-*NRT2* genes I identified the origins of the corresponding gene families as resulting from different rounds of duplications (chapter 3). The sub-functionalization of duplicated genes can result in the diversification of the expression following the duplication event according to the duplication-degeneration-complementation (DDC) model (Force et al., 1999). This differential regulation of gene expression has been suggested to be a common means of sub-functionalization and to happen very rapidly after the duplication (Wagner, 2000; Gu et al., 2002; Zhang, 2003). According to this theory, the expression regulation of different genes within the same gene family is hence indicative of their functional role.

In this chapter I investigate the environmental drivers of the N transporter clades modulation in order to validate and understand the different functional roles of the putative functional units previously designed. To do so I exploit the data provided by the *Tara* Oceans expedition, including all the metadata describing the environmental conditions of the sampling stations at the sampling time. To provide insights over diatom putative-functional units I assess in particular the optimal conditions for the expression of each clade.

5.3 Material and Methods

5.3.1 Data

Transporter abundances and normalization

The abundance of the unigenes selected by the phylogenetic analysis (chapter 3) was extracted by the metatranscriptomic and metagenomic dataset of *Tara* Oceans (Carradec et al., 2018). Occurrences values are computed as the fraction of the number of reads mapped per kb of unigene covered with reads per the total abundance of reads annotated to diatoms in the sample. The sum of occurrences for all the unigenes for a given sample is equal to 1. The dataset of di-*AMT1* and di-*NRT2* occurrences within the metatranscriptome dataset can be found respectively in Supplementary File 9 and 10 while in the metagenomic dataset the data refers to Supplementary File 11 and 12. The total number of reads sequenced per sample is available as well in Supplementary File 13, for the metatranscriptome and in Supplementary File 14 for the metagenome. The data was extracted by the cited public datasets by Dr. Eric Pelletier from Genoscope in Paris.

5.3.2 Data mining

The presence-absence of each clade for both families is defined by the presence in the metatranscriptome database or the metagenome database of *Tara* Oceans. A clades is considered present in a sampling site if the sum of the mRNA abundance of the unigenes belonging to the clade is higher than 0 in at least one of the four size-classes sampled (0.8-5 μm; 5-20 μm, 20-180 μm; 180-2,000 μm). A clade mRNA abundance is computed per each size class and per sample as the sum of the abundances of unigenes belonging to the same

clade. Whole gene family mRNA abundance are computed per each size class and per sample as the sum of the abundances of unigenes belonging to the same gene family.

5.3.3 Profiles of N transporters mRNA

To compare the mRNA levels of the two gene families, the transcripts abundances per family were summed per size class and then compared to the median of the obtained values per gene family. This comparison resulted in different classes of positive and negative divergence from the global median value computed on both sampling depths together. Focusing on the two gene families taken as a whole, the relationship with environmental parameters of the whole sum of transcripts per family was investigated through multiple Spearman correlations with the environmental variables available, with a *p*-value adjustment according to the Benjamini & Hochberg method (Benjamini and Hochberg, 1995). To be considered significant the correlation had to result in an adjusted *p*-value lower than 0.05 and an absolute value higher than 30. A further correlation was run between the mRNA levels of the two gene families at the different size classes and the number of reads annotated as diatoms in the corresponding metatranscriptomic samples.

5.3.4 di-*AMT1*and di-*NRT2* clades distribution

Zero-adjusted Bray-Curtis dissimilarity coefficients (size fraction 20-180 μm) were computed for the 106 *Tara* Oceans stations through the *vegan* R package (Oksanen et al., 2017). Stations have been clustered applying the Ward's minimum variance method (Murtagh and Legendre, 2014) and aggregated using a cutting level to form a number of clusters of stations. The clustering method choice as well as the optimal cutting level value were supported by the silhouette width (Rousseeuw, 1987) of the observations

developed through the R package *cluster* (Maechler et al., 2016). A number of 5 clusters of stations were defined for di-*AMT1* and di-*NRT2* clades over their mRNA abundance data.

5.3.5 Environmental PCA

A PCA was performed on all the sampling stations through the R package *FactoMineR* v.1.32 (Lê et al., 2008). A subset of environmental variables was selected specifically for each gene family as the subset of variables better explaining the difference between the communities using the *bioenv* function of the *vegan* package (Oksanen et al., 2017). This subset of variables is the one maximizing the Mantel correlation between the dissimilarity matrix of Euclidean distances between stations based on the environmental variables and the dissimilarity matrix of Bray-Curtis distances between the same stations based on the abundance of mRNA of the clades. Stations clades-based clusters were mapped on the PCA biplot through a discrete colorimetric scale.

5.3.6 Vertical switch

The zero-adjusted Bray-Curtis distance (Clarke et al., 2006) on the mRNA levels of di-*AMT1*and di-*NRT2* clades was computed to analyze the difference in composition of the communities at different depths. This value range from 0 to 1, where 0 means the communities of surface and at DCM are equal and 1 means they are completely different. This value has been confronted through a Spearman correlation to the environmental parameters of surface and of DCM, with a *p*-value adjustment of Benjamini & Hochberg (1995).The total sum of transcripts of di-*AMT1* and di-*NRT2* was computed and their ratio in surface over DCM was obtained per each station. A one-tail t-test tested the relationship between the two ratios to test if per *i* stations the following statement (Eq. 5.1) is true:

$$\frac{\Sigma AMT1_{SRF}}{\Sigma NRT2_{SRF\,i}} < \frac{\Sigma AMT1_{DCM}}{\Sigma NRT2_{DCM\,i}} \quad (5.1)$$

Moreover a Pearson correlation was computed between mRNA levels associated with N transporter clades and the relative mRNA levels associated to prokaryote N metabolism KOs (Sunagawa et al., 2015). Correlations were run separately at surface from DCM. The graphical representation of these correlations across the oceanic N cycle has been developed by Juan Jose Pierella Karlusich (ENS, Paris).

5.3.7 BRT model

To investigate which environmental variables have a role in the use of the clades a boosted regression tree (BRT) model (Elith et al., 2008) was run for each clade of the two gene families through the *dismo* and *gbm* R package (Ridgeway, 2006; Hijmans et al., 2017). This approach has been already described in chapter 2.3. The predictor variables used for the modeling were the same environmental variables previously selected (chapter 4.3.5). Models were run both over presence-absence data and mRNA abundances of the size fraction 20-180 μm using respectively Bernoulli and Laplace distributions. BRT was parameterized with slow learning rates (0.001-0.005) able to build models estimating reliable responses and tree complexity equal to 5, to include complex interactions. Models were run using a 50% bag fraction and a 10-fold cross-validation. For each clade the model was simplified identifying the variables that provided the best model performance. A k-fold cross-validation procedure was used to train (90%) and test (10%) each model and select the optimal number of trees. Statistical significance was assessed through cross validated AUC score (>0.7) for presence-absence based models (Tab. 5.3). The significance of mRNA-based models has been estimated through a coefficient of determination, expressed as the Pearson correlation coefficient between the observed and predicted mRNA levels for the same *Tara* Oceans stations,

classed in quartiles, applying then a threshold on the p-value (<0.05) and on the ρ ($|\rho| > 0.35$) (Tab. 5.4). To summarize the results of the BRTs model, the relative importance, or contribution, of each predictor variable was estimated. To better understand the relationship between clades and the environmental variables I built partial dependence plots of the major contributors of the models.

5.3.8 Sensitivity test

To have a first, rough idea of the possible impact of climate change on the diatom functionality, the machine-learning predictor was used for a sensitivity analysis to a change of the temperature. The probability of presence-absence of each clade in every sampling stations was predicted in different scenarios, with an increasingly higher variation in temperature (up to 3°C every 0.5°C), maintaining the other variables fixed to the observed values. The resulting probabilities were translated in presence absence applying a MaxSens+Spec threshold computed through the *PresenceAbsence* R package

Tab. 5.3: Statistics of BRT models run on presence absence data of clades.

Clade	Trees number	Mean tot. dev.	Mean res. dev.	Estimated c.v. dev.	Training data correlation	c.v. correlation	Training data AUC score	c.v. AUC score	Significant model
AMT1									
1	-								no
2	4550	0.38	0.198	0.279±0.086	0.652	0.481±0.147	0.978	-Inf	no
3	-								no
4	2950	0.781	0.629	0.725±0.072	0.48	0.269±0.14	0.871	0.641±0.107	no
5	400	1.174	0.794	1.012±0.061	0.657	0.414±0.104	0.898	0.786±0.049	yes
6	1900	1.174	1.151	1.168±0.014	0.531	0.14±0.061	0.848	0.617±0.039	no
7	450	0.535	0.245	0.363±0.073	0.744	0.603±0.147	0.978	0.875±0.069	yes
8	400	1.155	0.779	1.046±0.067	0.703	0.357±0.098	0.92	0.702±0.069	yes
9	500	1.377	0.739	0.933±0.146	0.747	0.645±0.094	0.926	0.828±0.055	yes
10	600	1.363	0.677	0.982±0.102	0.779	0.589±0.076	0.943	0.842±0.035	yes
11	700	0.487	0.203	0.38±0.09	0.799	0.504±0.142	0.984	0.829±0.075	yes
NRT2									
1	1000	0.781	0.772	0.774±0.047	0.465	0.111±0.108	0.853	0.622±0.088	no
2	-								no
3	800	1.092	0.523	0.901±0.087	0.797	0.426±0.097	0.971	0.807±0.046	yes
4	550	1.383	0.883	1.24±0.087	0.716	0.38±0.101	0.916	0.711±0.062	yes
5	3850	0.625	0.454	0.595±0.029	0.602	0.212±0.086	0.932	0.733±0.08	yes
6	750	1.135	0.592	0.87±0.104	0.738	0.516±0.082	0.938	0.829±0.048	yes
7	2550	1.357	1.112	1.318±0.041	0.611	0.198±0.098	0.858	0.601±0.054	no
8	350	0.911	0.587	0.749±0.063	0.632	0.417±0.094	0.921	0.857±0.038	yes
9	450	1.07	0.661	0.931±0.07	0.723	0.385±0.087	0.938	0.792±0.045	yes
10	700	0.487	0.182	0.256±0.067	0.752	0.789±0.105	0.986	0.964±0.024	yes
11	500	1.255	0.729	1.038±0.085	0.757	0.466±0.085	0.942	0.775±0.05	yes
12	2800	1.305	0.417	1.016±0.155	0.885	0.531±0.099	0.981	0.804+-0.048	yes

(Freeman and Moisen, 2008). From these results the ubiquity of each clade in every temperature scenario was calculated as the percentage of stations where the clade was present over the total number of stations sampled normalized over the observed scenario (temperature increase = 0°C).

5.4 Results and Discussion

Results are organized through four main sections to address the modulation of the two gene families at different levels. In the first section (chapter 5.4.1) the mRNA levels of the two gene families are analyzed as a whole, looking at the geographical patterns of their modulation and environmental linkages of these mRNA levels. In the second and third sections the mRNA levels of single clades are studied focusing on geographical- (i.e., horizontal, chapter 5.4.2) and along the water column- patterns (i.e., vertical, chapter 5.4.3) respectively. Finally, on the fourth section (chapter 5.4.4), a finer-scale analysis of the environmental modulation was run to understand the cues

Tab. 5.4: Statistics of BRT models run on expression data (20-180 μm) of clades.

Clade	Trees number	Mean tot. dev.	Mean res. dev.	Estimated c.v. dev.	Training data correlation	c.v. correlation	Pearson cor. RHO	Pearson cor. p-value	Significant model
AMT1									
1	50	1.70E+11	1.29E+11	1.29E+11±3.43E+10	0.352	0.368±0.118	0.51	3.05E-07	yes
2	50	3.03E+10	2.28E+10	2.36E+10±4.90E+9	0.283	0.32±0.126	0.56	1.01E-08	yes
3	50	2.79E+10	2.26E+10	2.36E+10±4.90E+09	0.284	0.361±0.074	0.47	2.71E-06	yes
4	50	2.59E+10	2.33E+10	2.43E+10±2.76E+09	0.406	0.27±0.11	0.51	3.95E-07	yes
5	50	8.58E+09	6.51E+09	6.62E+09±1.21E+09	0.42	0.442±0.095	0.61	2.36E-10	yes
6	50	1.94E+10	1.25E+10	1.24E+10±4.56E+09	0.305	0.341±0.081	0.61	2.43E-10	yes
7	-								no
8	50	2.67E+09	1.78E+09	1.82E+09±5.69E+08	0.403	0.34±0.099	0.58	3.60E-09	yes
9	50	4.76E+09	2.66E+09	2.69E+09±1.33E+09	0.334	0.617±0.12	0.69	5.05E-14	yes
10	50	3.22E+10	1.91E+10	1.90E+10±1.13E+10	0.228	0.452±0.093	0.6	5.44E-10	yes
11	-								no
NRT2									
1	50	1.83E+10	1.39E+10	1.43E+10±2.73E+09	0.494	0.524±0.108	0.52	1.78E-07	yes
2	50	1.13E+11	8.12E+10	8.14E+10±2.33E+10	0.293	0.438±0.117	0.7	4.80E-14	yes
3	50	1.49E+09	8.34E+08	8.10E+08±6.50E+08	0.208	0.652±NA	0.37	0.0003585	yes
4	50	8.35E+09	4.87E+09	4.95E+09±2.37E+09	0.309	0.253±0.093	0.56	2.26E-08	yes
5	50	7.42E+10	5.89E+10	6.09E+10±1.38+E10	0.194	0.207±0.098	0.42	4.91E-05	yes
6	50	7.67E+09	5.42E+09	5.47E+09±1.21E+09	0.442	0.348±0.108	0.62	8.53E-11	yes
7	50	8.05E+08	4.66E+08	4.79E+08±1.29E+08	0.111	0.118±0.18	0.19	0.0693	no
8	50	7.15E+08	3.87E+08	3.76E+08±2.40E+08	-0.034	-0.163±NA	0.25	0.01868	no
9	50	2.84E+09	1.61E+09	1.62E+09±7.86E+08	0.265	0.384±0.107	0.45	6.31E-05	yes
10	-								no
11	absent								no
12	50	3.53E+08	2.04E+08	2.07E+08±6.89E+07	0.496	0.174±0.153	0.45	1.13E-05	yes

behind N transporter clades distribution and abundances patterns thanks to a machine learning approach.

5.4.1 Whole gene family modulation

I studied the modulation of mRNA levels of di-*AMT1* and di-*NRT2* genes computing the deviance from the median abundance (see methods for further details). The two gene families show some correspondence in terms of geographical modulation but also several discrepancies (Fig. 5.1). From the literature we already know that diatoms cells differentially regulate the two gene families as response to external cues: e.g., *Thalassiosira pseudonana* downregulate *AMT1* and upregulate *NRT2* in N starvation conditions (Bender et al., 2014). The difference between the two biogeographies of mRNA deviance is thus mirroring the cardinal difference between the two regulatory systems, far more complex than a solely response to N sources availability.

Among the regions displaying similar deviances levels between the two gene families there are the Mediterranean Sea and the Indian Ocean where the mRNA abundances of N transporters are overall high, making an exception for the low abundances of the western Mediterranean basin. Di-*AMT1* mRNA levels are close or slightly higher than the median in the nitrate-rich tropical Pacific and Antarctic stations, whereas, di-*NRT2* mRNA abundance is low over the same regions. di-*AMT1* are abundant in areas of low N availability (e.g., Mediterranean Sea and Indian Ocean), otherwise they are not differentially modulated. Wherever di-*AMT1* transcripts are found abundant also di-*NRT2*'s are, with the exception of high N availability areas where di-*NRT2*s show low amount of mRNA. This same response is indicated also by the fact that the amount of di-*NRT2* mRNA shows negative correlations at both sampling depths to different sources of N such as NO_2^-, $NO_2^- + NO_3^-$, NO_3^-, NH_4^+. di-*NRT2* mRNA levels are correlated also with other major nutrients such as PO_4^{3-} and Si while it is positively linked with iron availability (Tab 5.5). The

same correlative exercise run on the sum of di-*AMT1* transcripts resulted in strong correlations to latitude and temperature, indicative of a large-scale regionalization of mRNA levels. di-*AMT1* mRNA levels from the size class 20-180 μm exhibit positive correlations to Si, suggestive of a relationship between the abundance of diatoms and di-*AMT1* mRNA levels (Tab 5.5). This same finding is supported by the correlations between the total mRNA annotated to diatoms and the di-*AMT1* and di-*NRT2* mRNA abundances (Fig. 5.2). It would appear that large diatoms cells (20-2,000 μm) present di-*AMT1* mRNA levels weighted over the whole-cell activity while there is no significant correlation together with small diatoms abundances or any size-fraction diatoms mRNA levels and the sum of di-*NRT2* transcripts.

5.4.2 Modulation at clade level – horizontal

We just saw how diatoms differently regulate *AMT1* or *NRT2* genes families but we know that the regulatory system behind is much finer than this. Different genes belonging to the same gene family have been detected to be differentially expressed. For example, transcriptomic studies on *P. tricornutum* found 4 *AMT1* genes upregulated and 2 *AMT1* genes downregulated in response to environmental cues like N starvation (Alipanah et al., 2015; Levitan et al., 2015), or in *P. multiseries* were found two *NRT2* genes upregulated and one *NRT2* gene downregulated facing similar stress conditions (Bender et al., 2014).

I expect thus di-*AMT1*and di-*NRT2* clades abundances to exhibit different modulations corresponding to different environmental conditions. Observing the distribution of mRNA across the different clades per station and per Oceanic basin I indeed find that they are differentially utilized according to the diatom size-fraction and to the global bioregions (Fig. 5.3).

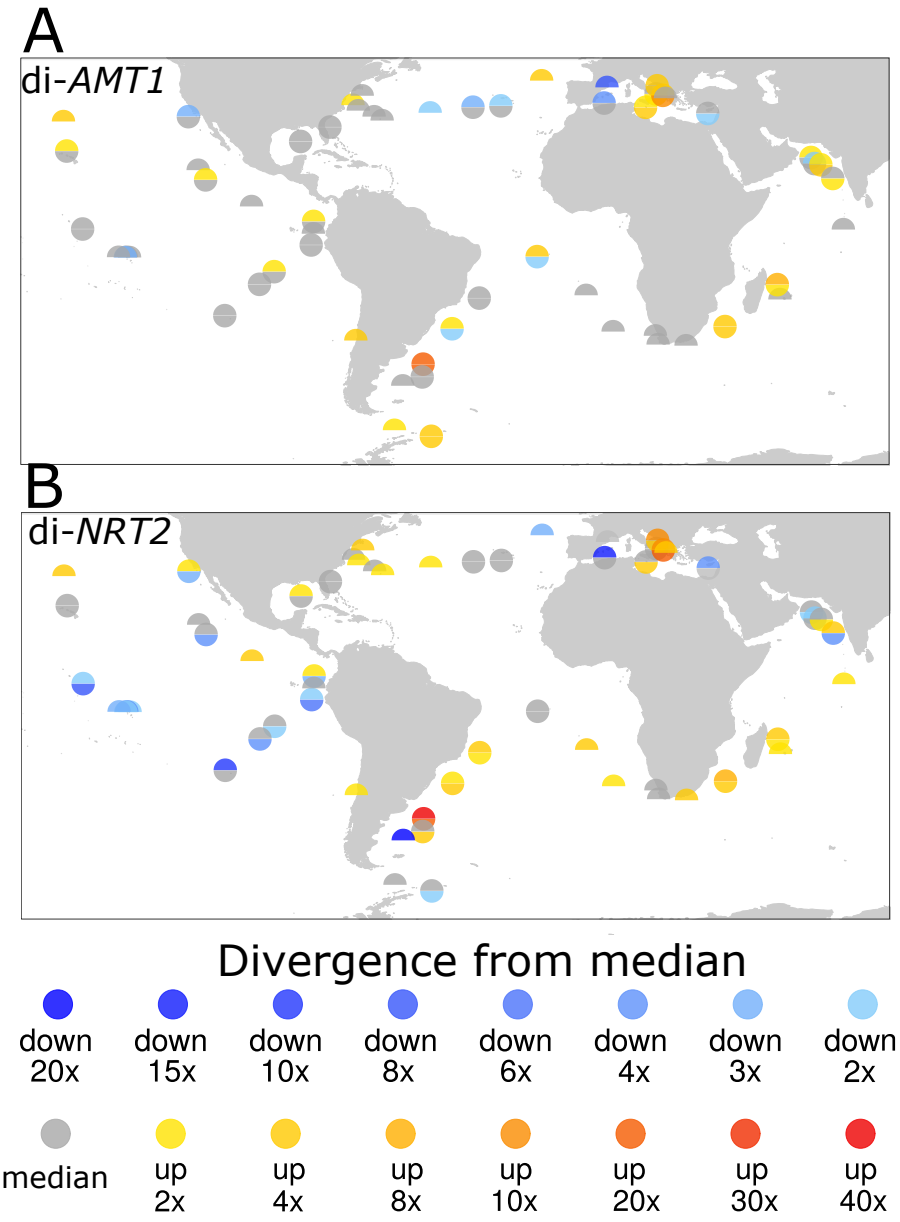

Fig. 5.1: mRNA levels of di-*AMT1* (A) and di-*NRT2* (B) at size class 20-180 μm. The sum of transcript assigned to each gene family present in every site is here expressed as fold-change over the median mRNA abundance value of the same gene family over the whole *Tara* Oceans dataset. The top portion of each circle represents samples collected at the surface and the bottom portion represents the DCM (stations missing metatranscriptome data for one of the two depths are drawn as half circles).

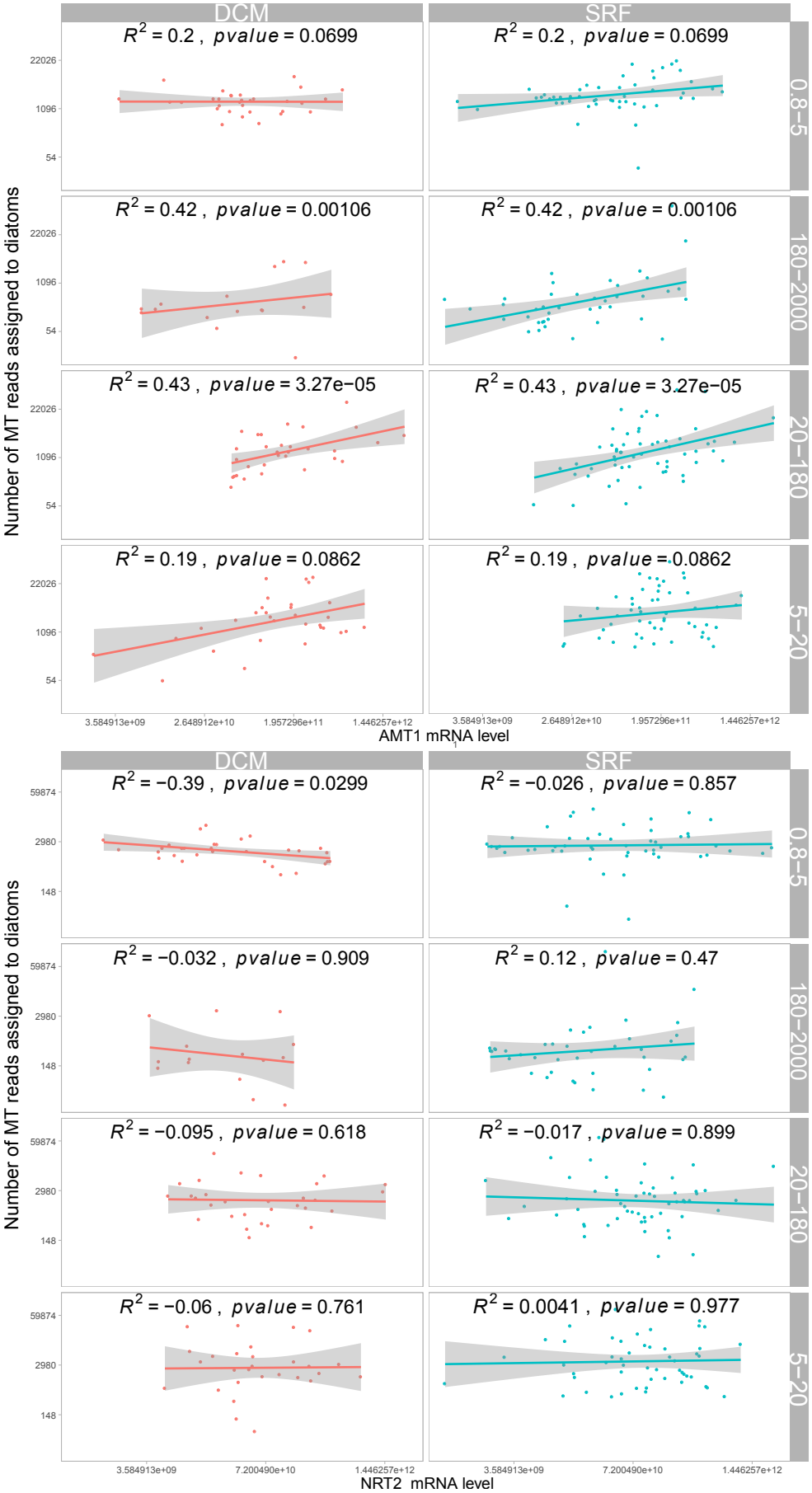

Fig. 5.2: Spearman correlations between the total mRNA abundance assigned to diatoms and mRNA levels of di-*AMT1* (A) and di-*NRT2* (B) respectively in 4 size classes.

Tab. 5.5: Spearman correlations between the sum of di-*NRT2* transcripts or the sum of di-*AMT1* transcripts and the environmental variables available in *Tara* Oceans for the 4 size classes of interest. Only the variables with a significant (adjusted p-value<0.05) correlation in surface or DCM are shown.

Size class	Environmental parameter	Spearman RHO	Adjusted p-value	Station depth
AMT1				
0.8-5	Mean Latitude	-0.38	0.0429	SRF
0.8-5	Mean Oxygen	0.53	0.0018	SRF
0.8-5	Mean Temperature	-0.41	0.0219	SRF
0.8-5	Oxygen Dissolved	0.49	0.0033	SRF
0.8-5	Silicate	0.38	0.0443	SRF
0.8-5	Temperature	-0.46	0.0081	SRF
20-180	Silicate	0.37	0.029	SRF
180-2000	Mean Density	0.72	0.0305	DCM
20-180	Fe (PISCES2)	0.64	0.0031	DCM
NRT2				
0.8-5	Fe (PISCES2)	0.39	0.0391	SRF
0.8-5	Fe (DARWIN-ECCO2)	0.52	0.0118	SRF
0.8-5	fqy2..phi_sat...	-0.43	0.0189	SRF
0.8-5	Lyapunov	0.41	0.0226	SRF
0.8-5	Mean.Depth.Max.O2	0.51	0.0032	SRF
0.8-5	Mean Longitude	0.45	0.0092	SRF
0.8-5	Mean Nitrates	-0.56	0.003	SRF
0.8-5	Nitrate	-0.46	0.0075	SRF
0.8-5	NO_2^-	-0.49	0.0046	SRF
0.8-5	NO_2^- (DARWIN)	-0.4	0.0253	SRF
0.8-5	$NO_2^-+NO_3^-$	-0.57	0.0004	SRF
0.8-5	NO_3^- (DARWIN)	-0.58	0.0002	SRF
0.8-5	Oxygen Saturation	0.39	0.0326	SRF
0.8-5	Oxygen Utilization	-0.43	0.015	SRF
0.8-5	PO_4^{3-}	-0.43	0.0158	SRF
0.8-5	Ratio NO_3^-_NH_4^+	-0.63	0.00002	SRF
0.8-5	Temperature Seasonality index	0.37	0.0467	SRF
20-180	Fe (PISCES2)	0.36	0.0382	SRF
20-180	Mean Nitrates	-0.48	0.0059	SRF
20-180	NH_4^+ (DARWIN)	-0.34	0.0485	SRF
20-180	Nitrate	-0.36	0.0378	SRF
20-180	NO_2^-	-0.4	0.0166	SRF
20-180	$NO_2^-+NO_3^-$	-0.53	0.0003	SRF
20-180	PO_4^{3-}	-0.45	0.0051	SRF
5-20	$NO_2^-+NO_3^-$	-0.46	0.0072	SRF
5-20	PO_4^{3-}	-0.37	0.0489	SRF
0.8-5	Mean Chloro	-0.67	0.0013	DCM
0.8-5	Nitrate	-0.5	0.0413	DCM
0.8-5	Phosphate (100 m)	-0.61	0.0211	DCM
0.8-5	Nitrates Seasonality index	-0.57	0.01366	DCM
20-180	Mean Chloro	-0.59	0.0083	DCM
20-180	Mean Density	0.5	0.0448	DCM
20-180	Mean.Depth.Max.O2	0.61	0.0077	DCM
20-180	Mean Nitrates	-0.59	0.01918	DCM
20-180	NO_2^-	-0.58	0.01099	DCM
20-180	$NO_2^-+NO_3^-$	-0.6	0.0075	DCM
20-180	Phosphate	-0.49	0.0441	DCM
20-180	Phosphate (100 m)	-0.58	0.0253	DCM
20-180	PO_4^{3-}	-0.62	0.0045	DCM
20-180	SI	-0.5	0.0389	DCM

In Figures 5.3A the relative mRNA abundance of di-*AMT1* clades is displayed for three different size fractions. Clade di-*AMT1*-11, which is specific to polar-centric-Mediophyceae, is overall very rare but occasionally very abundant only in the 0.8-5 μm fraction (stations 145 and 152 in the North Atlantic Ocean - NAO – station 68 in the South Atlantic Ocean - SAO). These results together with their taxonomic specificity may suggest evolutionary adaptation of this clade. Clade di-*AMT1*-10, ancestral to supergroup A, globally exhibits low mRNA abundances, making the exception of the Mediterranean Sea (MS) and in some specific stations (namely, station 36 – Indian Ocean (IO) and station 92 – South Pacific Ocean (SPO)) where it is relatively abundant.

Also, clades included in the supergroup A exhibit specific patterns. They show higher mRNA levels at global scale, regardless of the size fraction (i.e., clades di-*AMT1*-1, di-*AMT1*-4 and di-*AMT1*-5 in fraction 20-180 μm, clades di-*AMT1*-2 and di-*AMT1*-4 in fraction 5-20 μm). Clade di-*AMT1*-3 in particular, which notably includes araphid pennate diatoms, is abundant in the size fraction 0.8-5 μm: it can be detected across all the oceanic regions but mostly in the very oligotrophic Indian Ocean (IO) where it can be defined the dominant clade. From a biogeographic point of view, the Mediterranean basin and the Antarctic area were the two regions most differentiated. Concerning the peculiarity of the Mediterranean basin, clades di-*AMT1*-2 and di-*AMT1*-10 are mostly exclusively found in this region while clades di-*AMT1*-4 and di-*AMT1*-5 are present everywhere with the exception of this same area. Instead, in the Antarctic stations, even if only few samples are available, they are all characterized by a high mRNA abundance of clade di-*AMT1*-6 and the peculiar presence of di-*AMT1* clades 7 and 9, found almost exclusively in these stations.

The same graphical representation of mRNA abundances for the di-*NRT2* gene family is represented in Fig. 5.3B. The basal clade di-*NRT2*-12 is relevant only in two stations located in NAO (station 4, fraction 0.8-5 μm) and in IO

(station 52, 5-20 μm). The only two *NRT2*-clades belonging to all the four main diatom groups (clade di-*NRT2*-2 and clade di-*NRT2*-6) unexpectedly do not always show high mRNA levels. Indeed, clade di-*NRT2*-2 is dominant in the MS and in some stations of the SPO, while clade di-*NRT2*-6 is relatively abundant in the Atlantic Ocean. The supergroup A basal clade di-*NRT2*-11 is specifically important in the Mediterranean Sea, the South Pacific Ocean and also in West North Atlantic Ocean of small diatoms (0.8-5 μm). Antarctic stations are peculiarly characterized by clades di-*NRT2*-6 and di-*NRT2*-10, which are mostly confined to this extreme region.

Several size class effects on the geographical distribution of di-*AMT1*and di-*NRT2* clades emerged from the above analyses. In particular, size fraction 0.8-5 μm diverged from larger diatoms profiles in terms of clade specific mRNA levels. Given the different ecological strategies of large and small diatoms toward nutrient uptake (Marañón et al., 2013), it is not unexpected for 0.8-5 μm size class diatoms to deploy a partially different set of N transporters. Relatively to their volume, smaller cells perform indeed a lower N uptake explained by lower storage capacities (Marañón et al., 2013). This may be due to a decreasing density of transporters proteins compared to larger cells (Aksnes and Cao, 2011) but I may speculate that also the presence of different evolutionary solutions for N transporters may be the cause of such different N uptake abilities.

To study the biogeography of N transporter clades I run a similar exercise to the one performed in chapter 4.4.2. The clustering of stations based on the relative mRNA levels in the 20-180 μm fraction (Fig. 5.4 and 5.5) grouped stations differently from the clustering based on the presence-absence data, with less geographical discrimination. Indeed, while for the presence-absence based clustering we could actually detect geographical patterns, the mRNA levels-based clustering is rather explained by the consequent environmental PCA. The distribution of clusters finds N sources availability (both NO_2^- and

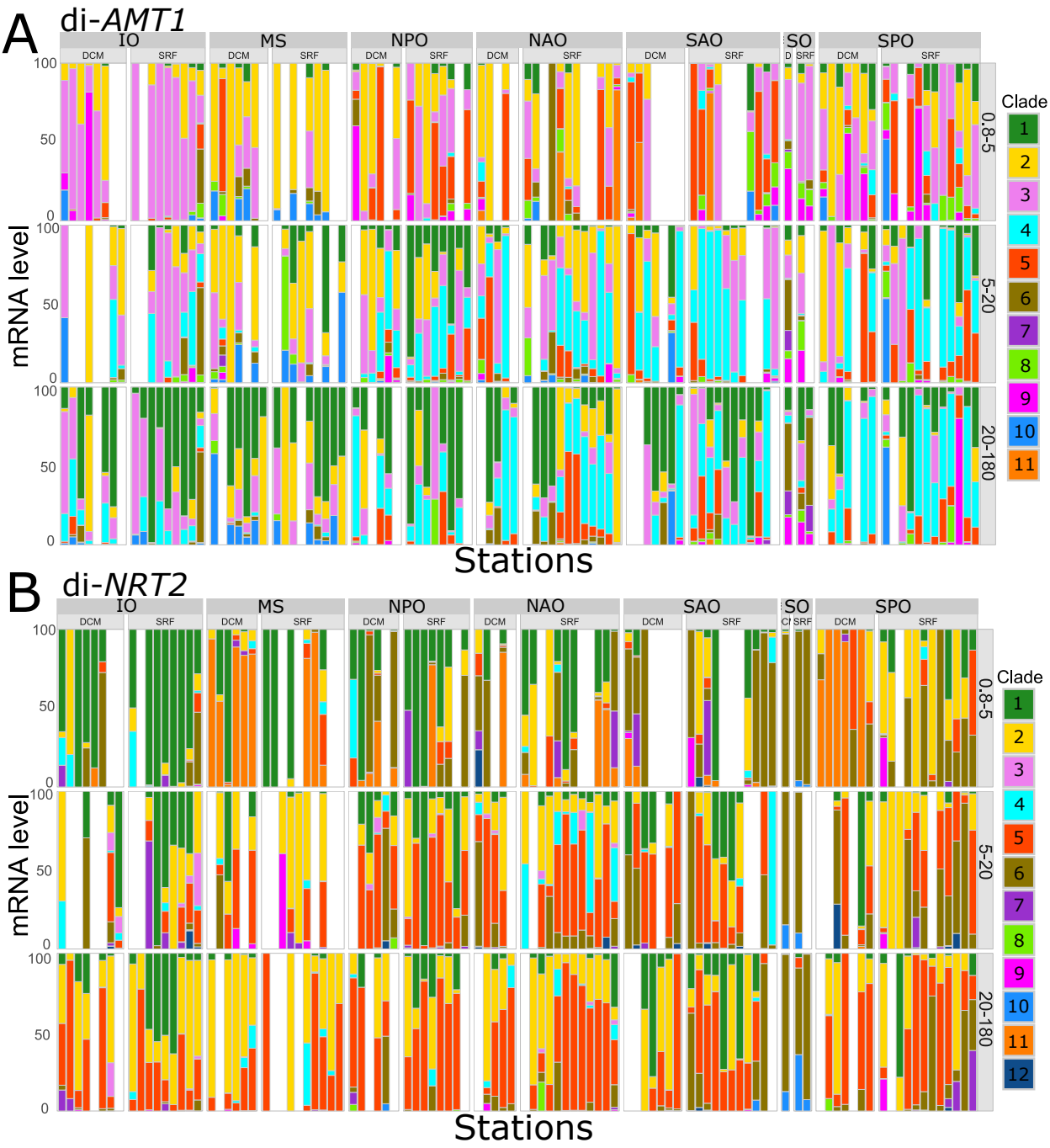

Fig. 5.3: Barplot of di-*AMT1* (a) and di-*NRT2* (B) clades relative expression in three size classes (0.8-5 μm; 20-180 μm; 5-20 μm). Sampling stations are clustered according to the sampling depth, surface (SRF) or DCM and oceanic basin: IO: Indian Ocean; MS: Mediterranean Sea, NPO: North Pacific Ocean; NAO: North Atlantic Ocean; SAO: Southern Atlantic Ocean; SPO: Southern Pacific Ocean and SO: Southern Ocean.

$NO_2^- + NO_3^-$) being the explicative variable of both families (Tab 5.6). Interestingly, most clusters of stations are strongly dominated in terms of mRNA abundances by one or two clades over the others (Fig. 5.4). Precisely, mRNA levels of clades di-*AMT1*-4, di-*AMT1*-6 or di-*AMT1*-9 and di-*NRT2*-6 dominate the clusters di-*AMT1-cyan*, di-*AMT1-pink* and di-*NRT2-pink*, respectively. All these clusters are generally characterised by high N availability (Fig. 5.5). Other di-*AMT1* clusters generally occupy low nitrate availability regions, not better defined by the PCA I performed. By contrast, for di-*NRT2* we can detect clusters characterized by high iron availability (cluster di-*NRT2-red*, dominated by di-*NRT2*-2), low nitrate and iron but high nitrite availability (clusters di-*NRT2-yellow*, di-*NRT2-cyan*, dominated by di-*NRT2*-5 the first and by di-*NRT2*-5 and di-*NRT2*-2 the second) and tropical, oligotrophic stations (cluster di-*NRT2-blue*, dominated by di-*NRT2*-1).

The result of pairwise correlations between environmental parameters and the mRNA abundances of clades revealed that overall di-*AMT1* and di-*NRT2* clades show negative correlations with N related variables (Fig. 5.6). This may indicate that in high N availability condition diatoms mostly respond by decreasing the abundance of N uptake mRNA. This is not unexpected considering the fact that most *AMT1* and *NRT2* genes have been found upregulated in case of N starvation conditions (Tab. 5.1 and 5.2 and references therein). However, a clear exception is given by di-*NRT2* clade 6 which is strongly positively correlated with different sources of N availability. This may indicate a striking different regulation of this particular evolutionary clade to be used for N storage. N storage is a strategy mostly adopted by large diatoms thanks to the presence of internal vacuoles for nutrient storage (e.g., Antia et al., 1963). These particular structures can represent 30%–90% of total cell volume in diatoms larger than 5 μm (Smayda, 1970). Storage capacity is usually larger in larger cells: through this strategy they fill up more slowly and they can longer sustain high uptake rates (Verdy et al., 2009). This strategy is optimal

for cells to adapt to environmental conditions characterized by intermittent high nutrient availabilities (Falkowski and Oliver, 2007).

5.4.3 Modulation at clade level – vertical

In stratified conditions, the nutricline acts as a boundary between two different fates of ammonium. The phytoplankton uptake and microbial oxidation are the two processes involving NH_4^+ at surface and at the DCM respectively. It is the differential availability of ammonium or nitrate which determines phytoplankton community structure along the water column (Wan et al., 2018). Indeed, according to the view of Wan et al. (2018), above the nutricline the phytoplankton communities are usually dominated by prokaryotes, having higher affinity for NH_4^+, while below the same boundary the abundant NO_3^- lead to the dominance of eukaryotic phytoplanktons which are competitive for the use of this N source.

Tab. 5.6: Bioenv output for the selection of the environmental parameters for the Principal Component analyses on relative mRNA abundances values.

Environmental parameters subset	**Correlation**
di-*AMT1*	
NO_2^-+NO_3^-	0.2845
NO_2^-+NO_3^-, NO_2^-	0.2934
Mean_Chloro,NO_2^-+NO_3^-, NO_2^-	**0.3131**
Mean_Chloro, NO_2^-+NO_3^-, NO_2^-, Temperature	0.2885
Mean_Chloro, Fe, NO_2^-+NO_3^-, NO_2^-, Temperature	0.2790
Mean_Chloro, Fe, Monthly_ipar, NO_2^-+NO_3^-, NO_2^-, Temperature	0.2629
Mean_Chloro, Fe, Monthly_ipar, NO_2^-+NO_3^-, NO_2^-, Temperature, Nitrocline	0.2550
Mean_Chloro, Fe, Monthly_ipar, NH_4^+, NO_2^-+NO_3^-, NO_2^-, Temperature, Nitrocline	0.2356
di-*NRT2*	
NO_2^-+NO_3^-	0.2624
Monthly_ipar, NO_2^-+NO_3^-	0.2759
Monthly_ipar, NO_2^-+NO_3^-, Temperature	0.2817
Fe, Monthly_ipar, NO_2^-+NO_3^-, Temperature	0.2809
Fe, Monthly_ipar, NO_2^-+NO_3^-, NO_2^-, Temperature	**0.2819**
Fe, Monthly_ipar, NH_4^+, NO_2^-+NO_3^-, NO_2^-, Temperature	0.2803
Mean_Chloro, Fe, Monthly_ipar, NH_4^+, NO_2^-+NO_3^-, NO_2^-, Temperature	0.2702
Mean_Chloro, Fe, Monthly_ipar, NH_4^+, NO_2^-+NO_3^-, NO_2^-, Temperature, Nitrocline	0.2524

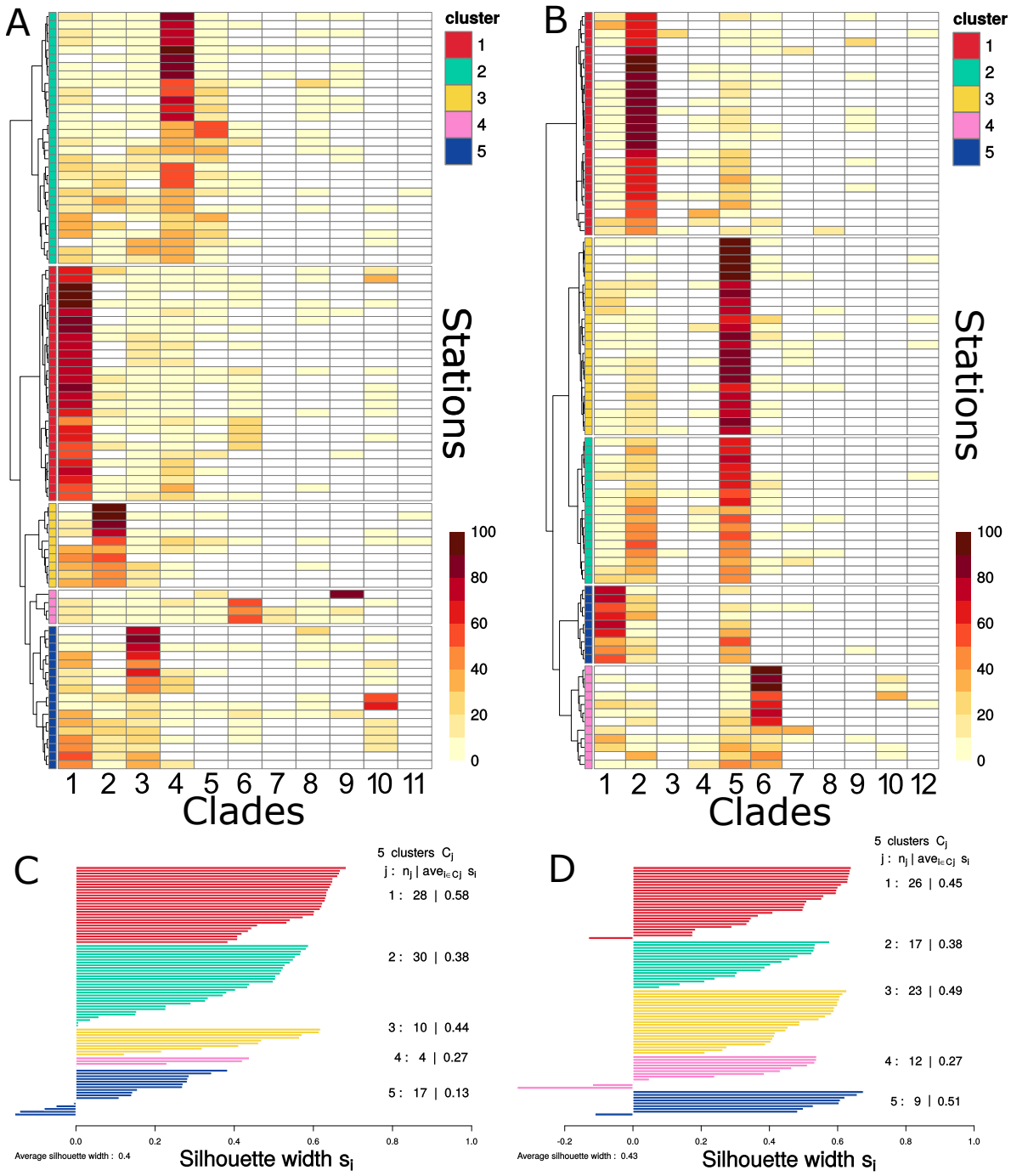

Fig. 5.4: Heatmaps showing di-*AMT1* (a) and di-*NRT2* (B) clades relative mRNA levels (20-180 μm) in the *Tara* Oceans stations. Stations are clustered by a Ward clustering method based on zero-adjusted Sørensen dissimilarity and annotated in 5 resulting clusters for di-*AMT1* and 5 clusters for di-*NRT2*. The white cells correspond to the stations where the corresponding clade is absent; the colored palette indicates the relative mRNA level of the clade. The normalization of the mRNA levels is built to grant a total normalized mRNA level per station up to 100. In panels C and D are the estimated silhouette values for each established cluster of di-*AMT1* and di-*NRT2*. Values closer to 1 indicate a high degree of similarity of the station within the cluster, positive values close to zero indicate stations which are closer to the other clusters, while negative values indicate stations which may have been misplaced by the clustering.

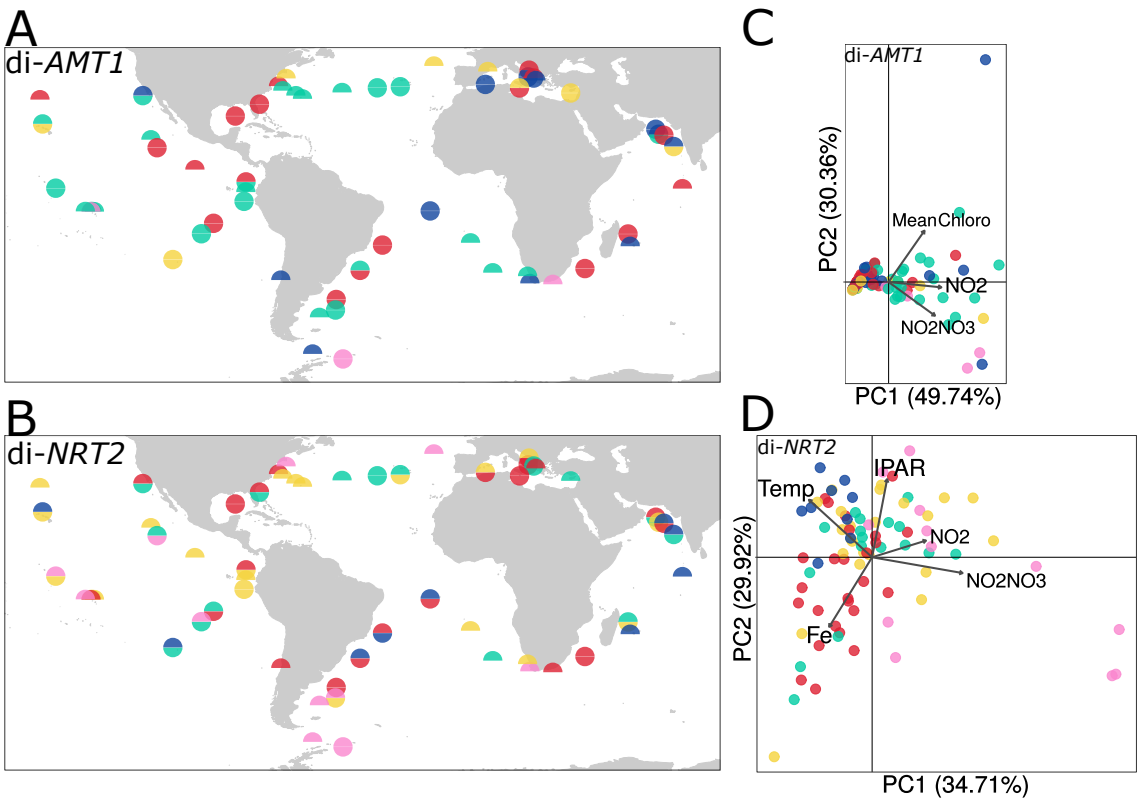

Fig. 5.5: Geographical clusters on di-*AMT1* (A) and di-*NRT2* (B) clades mRNA levels. The top portion of each circle represents samples collected at the surface and the bottom portion represents the deep chlorophyll maximum (stations missing metatranscriptome data for one of the two depths are drawn as half circles). Biplots of the environmental PCA of di-*AMT1* (C) and di-*NRT2* (D). Each point corresponds to a sampled station colored according to the cluster it belongs to, while the arrows correspond to the descriptors of the PCA space. Clustering of stations is based on the relative mRNA level of clades. On both gene families station clustering resulted in 5 clusters. Clusters are identified by the colors: *yellow*, *cyan*, *pink*, *blue* and *red*, and are defined in Fig. 5.4 (see methods).

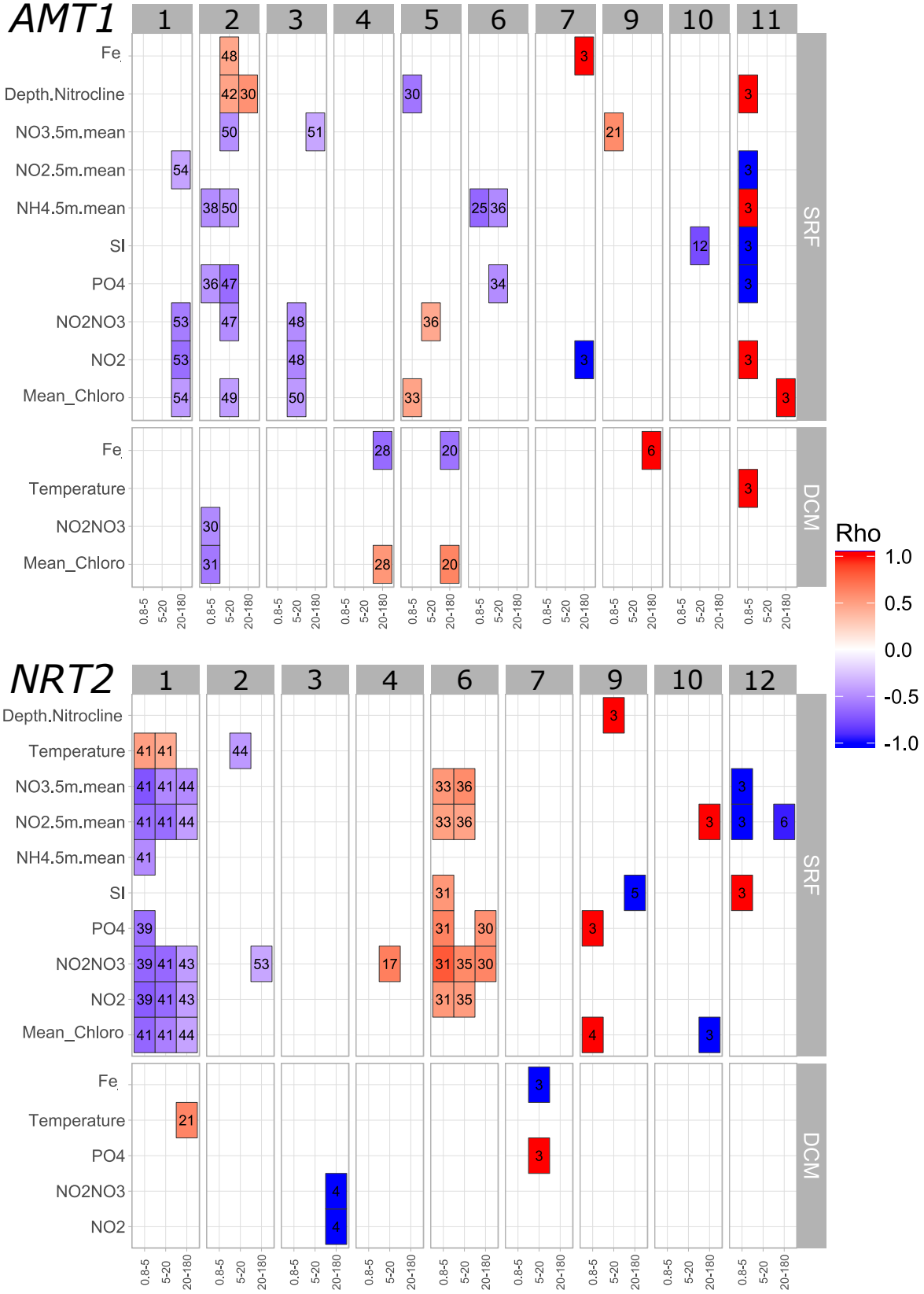

Fig. 5.6: Pairwise Spearman correlation between N transporter clade mRNA level and environmental parameters. Correlations are run for every size class and fdr adjusted. Only significant correlations are shown (p-value <0.05), and the number of samples on which the correlation is computed is written in the corresponding cell.

To investigate whether the different nutrient availability is also mirrored by a differential use of diatoms di-*AMT1* and di-*NRT2* I correlated the dissimilarity distances between the two sampling depths of every station to the environmental parameters measured in surface or at the DCM depth (Tab 5.7). Diatoms at the DCM should be taxonomically equivalent to the one found at surface, however their cell activity, and thus their use of N transporters is expected to dramatically change in relation to the strong change of conditions (starting by the differences in light and nutrients availability). A structural change on the compositions of clade expressed should reflect the higher nutrient availability and the lower light found at the DCM. The gradient along the water column of di-*AMT1* mRNA abundances was found to be strictly linked to light and NH_4^+ availability in surface and seasonal nitrate availability at the DCM. By contrast, only the carbon flux measured at 150 m was found to significantly explain the gradient in mRNA abundances across di-*NRT2* clades. Carbon fluxes strongly depend on the CO_2 uptake of phytoplankton and the consequent particulate organic carbon (POC) sinking (Guidi et al., 2015). The higher POC sinking, corresponds to specific communities at surface (Guidi et al., 2015), thus to different levels of competition in surface for N sources but also a higher amount of ammonium remineralized at depth. I may suppose that the strongest the carbon flux, the strongest are the differences in nitrogen availability of the two N forms. I hypothesize that the differential use of di-*NRT2* clades along the water column may be a direct consequence of the different sources of N availables.

To investigate the differential use of *AMT1* rather than *NRT2* and, thus, the differential N sources preferences of diatoms according to the depth, I computed the ratio of the total di-*AMT1* mRNA over the total di-*NRT2* mRNA at the two sampling depths (Fig. 5.7). Results show that in small diatoms (0.8-5 μm) the di-*AMT1* di-*NRT2* ratio is very low in oligotrophic regions whereas it is very high in regions with relatively high nutrient availability. This bimodal pattern is partially lost in larger diatoms suggesting again a size-class effect in

N uptake in diatoms (Li, 2002; Marañón, 2015; Van Oostende et al., 2017). While small diatoms may strongly respond to different N sources availability, larger cells do not show strong differences among different environmental conditions. I could speculate that being the surface area to volume ratio higher in smaller cells they may have a higher ability in controlling the uptake and rapidly respond to environmental conditions. The other possible explanation is that in oligotrophic regions small diatoms are actually thriving while larger cells are either using symbiosis for getting nitrate or in a low metabolic state, waiting for pulses of nutrients by eddies or storms.

The di-*AMT1* di-*NRT2* ratio was significantly lower at surface than at the DCM depth (p-value = $4.2e^{-12}$; Fig. 5.8) revealing that, generally, di-*AMT1* mRNA is relatively more abundant than di-*NRT2* mRNA at the DCM compared to surface. This may be the consequence of an actual expression response of diatoms to the availability of the two forms of N (with NO_3^- more abundant at DCM, generally) and thus it is the emerging pattern that one should expect if the Wan et al. (2018) scenario is generally valid.

Finally, to have a better understanding of the differential role of clades along the water-column I correlated the mRNA clade abundances with prokaryote N metabolism modules from the two sampling depths (Fig. 5.9). Generally, a correlation between the two would be expected in reason of the public goods: the more the prokaryotes fix and remineralize nitrate, the more ammonium is available to diatoms and thus the more ammonium transporters genes should be expressed to retrieve this pool of N. The comparison of the resulting correlation matrices (Fig 5.9) shows that a higher number of modules are found significantly correlated to clades at surface in respect to DCM. This result suggests a likely tighter compartmentalization between diatoms and prokaryotes N utilization at surface than at DCM, which is an important observation, suggesting that at DCM diatoms tend to use NO_3^- diffused from below and not remineralized locally while at surface they seems to profit

of the prokaryote activity. Not surprisingly, very few matches are coherent between the two sampling depths: they are all related to di-*AMT1* clades and linked to prokaryotic processes producing ammonium, such as N_2 fixation, assimilatory nitrate reduction to ammonium (ANRA) and dissimilatory nitrate reduction to ammonium (DNRA). The fact that di-*AMT1* clades have a clear relationship with NO_2^- releasing processes is coherent with what already observed in chapter 4 that this remineralized source of N is strongly informative of the biogeography of communities designed over this gene family. This depth-independent behavior may be justified by the consequential use of the public goods, i.e., where a high amount of prokaryotes produce ammonium, transporter clades for the uptake of the same substrate are highly expressed (di-*AMT1*-4, di-*AMT1*-8 and di-*AMT1*-9). Another supportive result of this latter hypothesis is that most of the correlations are located at surface, where phytoplankton is expected to have higher affinity for NH_4^+ in response to nitrifiers' ammonium recycling activity (Wan et al., 2018).

The most important finding of this exercise is that significant correlations are found and that they are mostly positive. This means that when the bacterial N-cycling activity is higher, the di-*AMT1* are also more active. In turn, this should correspond to a lower availability of NH_4^+. This could be interpreted as a higher priority for diatoms to use ammonium in oligotrophic conditions. A significant correlation between a prokaryote module and a N transporter clade indeed may mirror the consequent use of public goods as previously stipulated but also a similar environmental based regulation. It is to remind that the results of correlations are difficult to interpret because 'correlation is far easier to demonstrate than causation' (Sugihara et al., 2012).

5.4.4 Niche exercise

The pipeline of analysis I developed through this thesis to study diatoms functional diversity (chapters 3, 4 and 5) is based over the working hypothesis

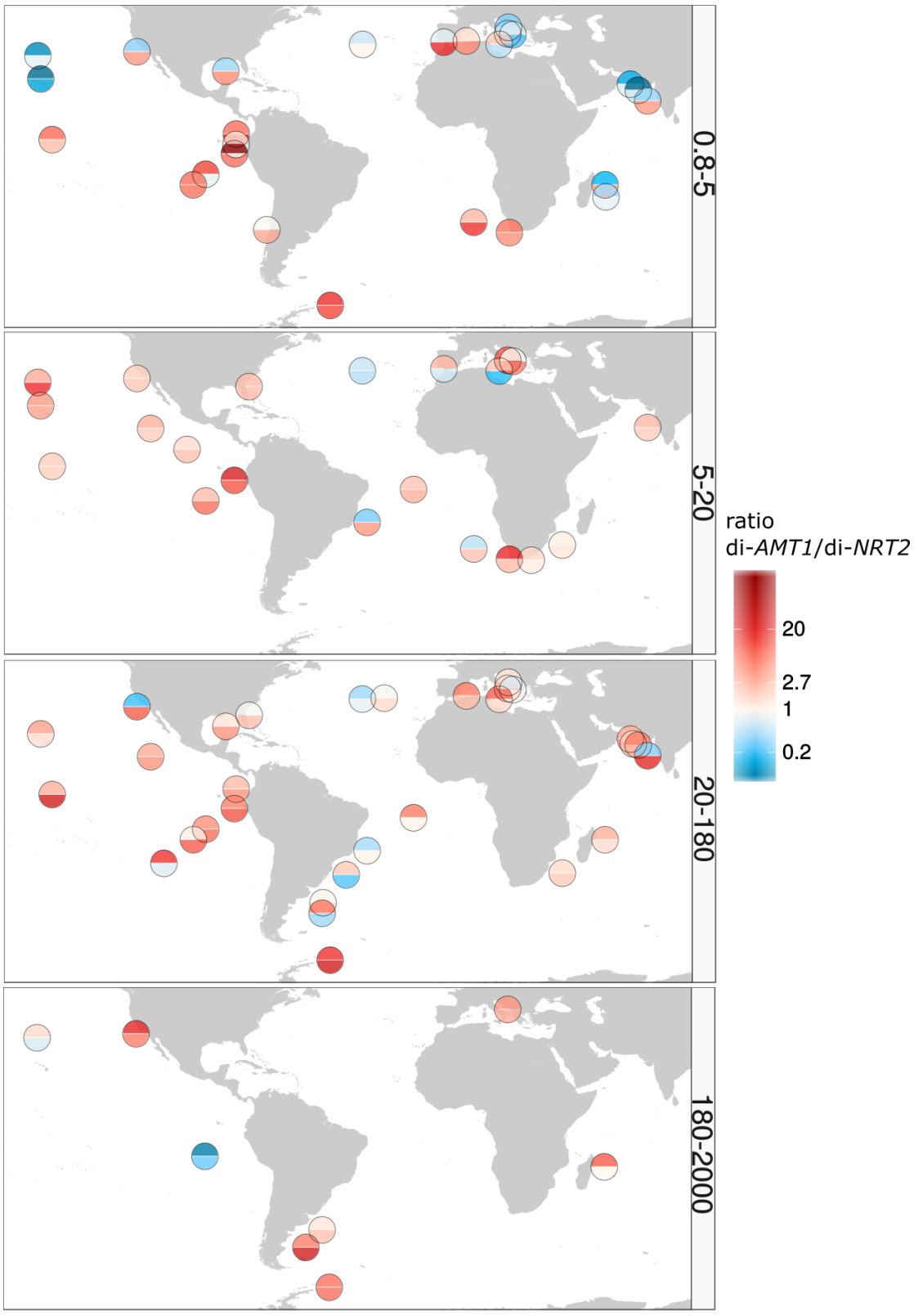

Fig. 5.7: Ratio between di-*AMT1* total mRNA abundance and di-*NRT2* total mRNA level computed in surface (top portion of each circle) and at DCM (down portion of each circle). Every panel refers to one of the four size classes taken into account.

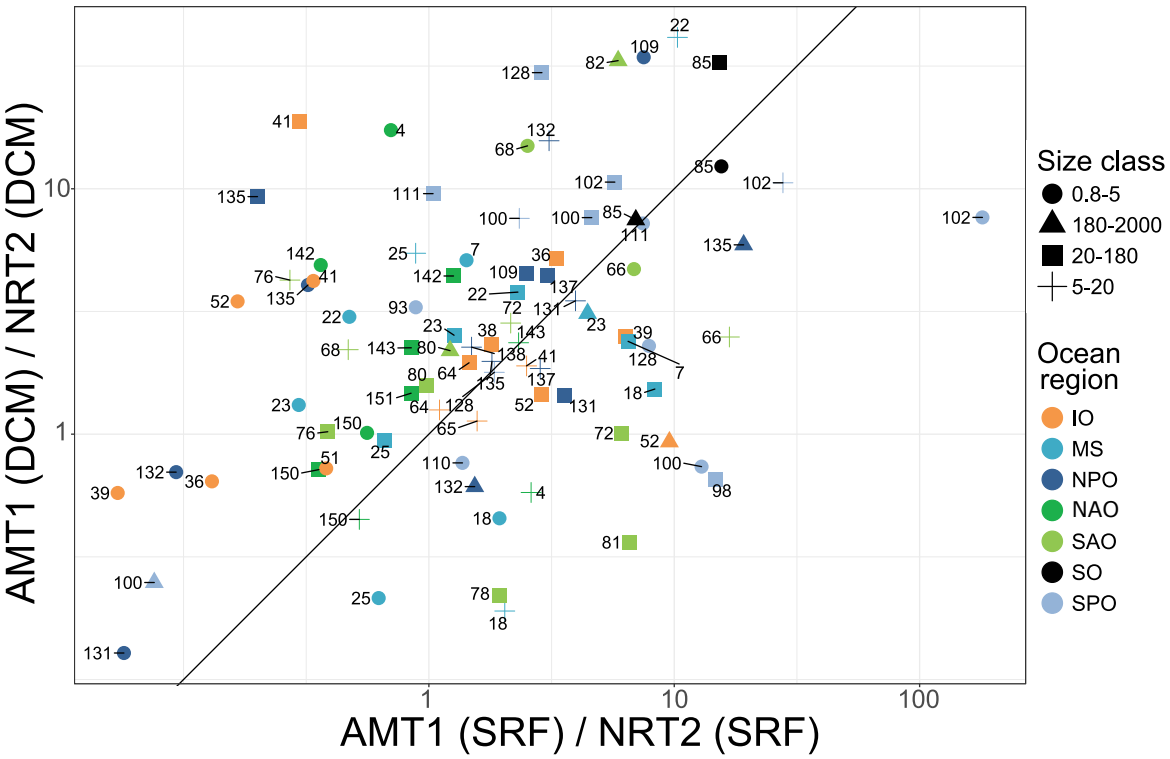

Fig. 5.8: Scatterplot of the ratio between di-*AMT1* total mRNA level and di-*NRT2* total mRNA level computed in surface (x axis) and at DCM (y axis). Stations are labeled according to the *Tara* Oceans station number, shaped according to the size classes they refer to and colored according to the ocean basin where they are located. In this scatterplot only the sampling which had mRNA values higher than zero of at least one di-*AMT1* gene and one di-*NRT2* gene both in surface and DCM are included. The distribution of point is significantly over the intercept, indicating that the ratio at DCM is significantly higher than at surface.

Tab. 5.7: Spearman correlations between the zero-adjusted Bray-Curtis distance between surface and DCM samples of the same station and the environmental variables available in *Tara* Oceans. Only the correlations with a significant (p-value<0.05) correlation are shown.

Size class	Gene Family	Environmental parameter	Depth of env. data	Spearman RHO	Adjusted p-value
0.8-5	di-*AMT1*	Nitrocline depth	SRF	-0.6075	0.030
20-180	di-*AMT1*	Monthly_ipar	SRF	0.5119	0.027
20-180	di-*AMT1*	NH_4^+	SRF	0.5363	0.016
0.8-5	di-*AMT1*	Mean Angular Scatt. coeff. (117-470 nm)	DCM	0.5613	0.035
5-20	di-*NRT2*	Mean Flux 150	DCM	-0.6678	0.049

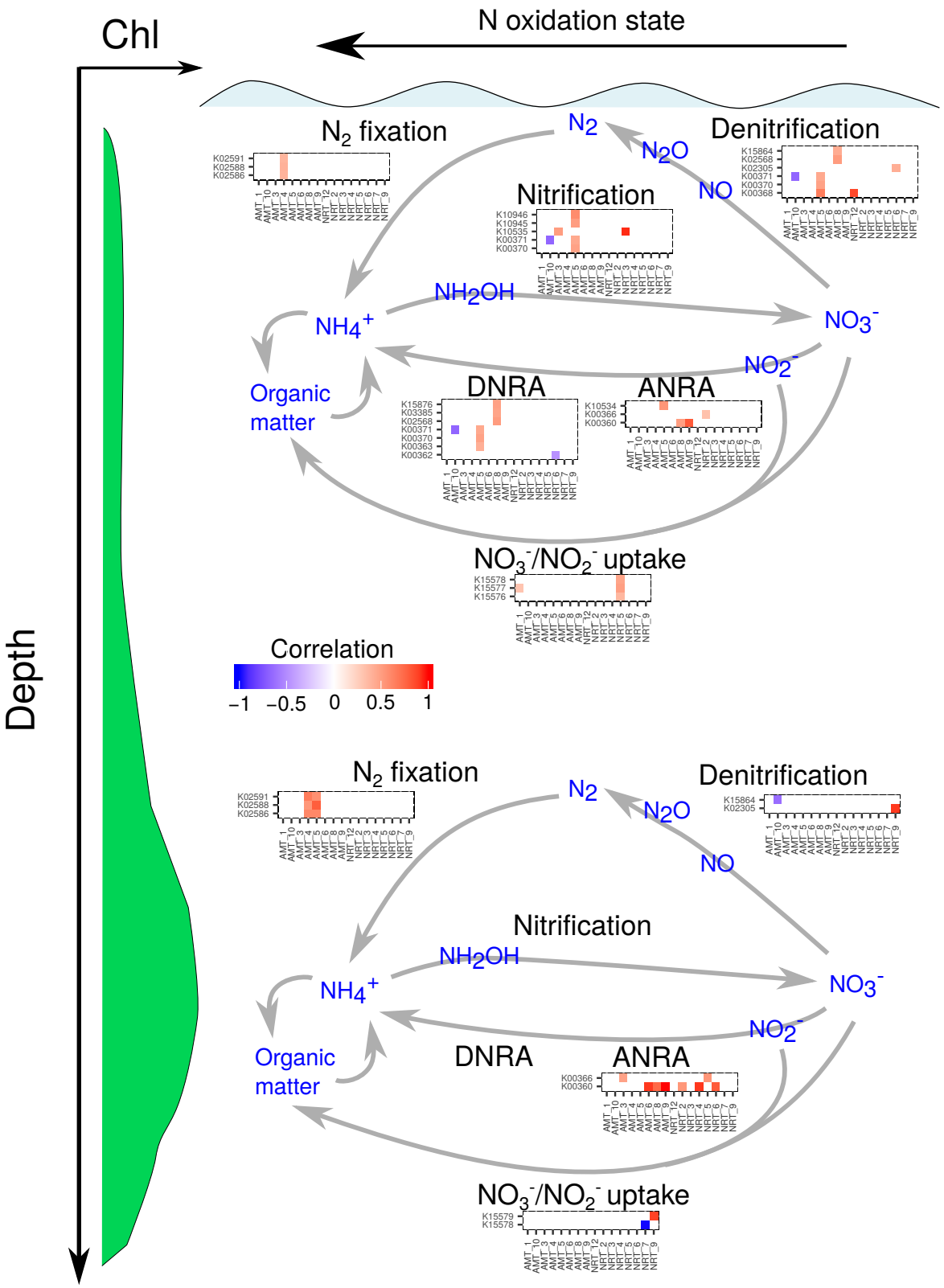

Fig. 5.9: Correlations of diatom di-*AMT1* di-*NRT2* clade mRNA abundances from 20-180 μm size fraction against prokaryotic nitrogen metabolism gene abundance from 0.22-1.6/3 μm size fractions. Abbreviations: DNRA, dissimilatory nitrate reduction to ammonium; ANRA, assimilatory nitrate reduction to ammonium.

that N transporter evolutionary clades may have a functional role which can be used as describer of the N utilization functional trait for diatoms. This information has a double level: its presence and its abundance. Indeed, one information is if diatoms own a particular adapted clade and one other is how they make use, through modulation, of this tool. To conclude this putative-functional study of diatoms across the global ocean I hereby chose a machine learning approach to detect which are the environmental drivers of clade distribution (presence-absence) and modulation (mRNA levels). This ecological niche derived tools is named boosted regression trees (BRT). Among the several reasons behind the choice of this tool is its ability to fit complex non-linear relationships, the possibility to apply any type of prediction variables (numeric, categorical, binary, etc) and to use not only binary response variables but also numeric ones, and finally to admit missing values in the prediction variables (Elith et al., 2008).

Through this tool I was able to weight the role (contribution) of nine abiotic and biotic parameters in determining the presence and the mRNA levels of N transporters clades. For every clade two models were developed using presence-absence and mRNA abundances at 20-180 μm size class respectively. Of note, an important limit of this approach happened to be the ubiquity of clades: it was not possible to model the distribution of ubiquitous clades, as they have no environmental preferences, but I met also difficulties in modeling rare clades mRNA levels as the number of occurrences were too low. Concerning the ubiquitous clades, the presence-absence modeling is no needed as no environmental limits are encountered by these clades. For rare clades the mRNA levels modeling may be replaced by the presence-absence one instead. This may be supported by the speculation that the use of such clades is limited to environmental conditions so rare that the model for their distribution could be indicative of their modulation as well.

I was able to model 17 (over 22) clades mRNA abundances and 15 (over 23) clades presence-absence (Fig. 5.10). The major contributors of the models are iron, nitrate, chlorophyll α and temperature for both di-*AMT1* and di-*NRT2*, while the latter is not very important for di-*AMT1* (Fig. 5.10). Concerning mRNA abundances iron and NO_2^-+NO_3^- are by far the most important contributors for di-*AMT1* transporters while di-*NRT2* mRNA levels are also controlled by the nitrocline depth and NO_2^-.

Of the above-mentioned variables, iron and temperature show the most marked large-scale geographical gradients (Fung et al., 2000; Hansen et al., 2006). The distribution of N transporter is strongly explained by these variables and this is suggestive at some extent, of a distribution, in particular of di-*NRT2*, following geographical gradients at basin and latitudinal scales. Nevertheless, the mRNA abundance-based models weight the relative contribution of N up to values close to 100%, indicating the mRNA level of specific clades to be in some cases controlled by the availability of their substrate.

We observe a redistribution of the contribution of variables between mRNA levels and presence-absence models that is slightly different according to the gene family (Fig. 5.10 and 5.11). Focusing on the two major contributors among the explicative variables, iron and NO_2^-+NO_3^- availability, it is clear the difference between the environmental cues used for distribution and modulation (Fig. 5.11). Clade mRNA level is relatively more controlled by N availability rather than the corresponding models for the presence of most of di-*NRT2* and selected di-*AMT1* (Fig. 5.11). By contrast, the use of iron availability as descriptor finds very similar median values for the two data-based set of models, but the wider variance in mRNA-based models suggests a more differentiated use of this signal for gene expression rather than distribution (Fig. 5.11). Of note, in di-*AMT1* there is a higher contribution of the recycled form of nitrogen, NO_2^- (as for di-*AMT1*-1, di-*AMT1*-3 and di-*AMT1*-6) while others are mostly explained by nitrate availability alone

(di-*AMT1*-5, di-*AMT1*-8 and di-*AMT1*-9), suggesting that mRNA levels are herein modulated by the availability of different sources of nitrogen.

Unexpectedly, di-*AMT1* clade 11 shows no sign of environment-related specificity (Fig 5.10). It is thus possible that the emergence of this ancestrally diverging polar-centric-Mediophyceae clade responded to the need for increased functional redundancy in given species. The mRNA levels of clade 10 di-*AMT1*, basal to supergroup A, is strongly influenced by iron concentration, with higher abundances at higher iron availabilities (Fig. 5.12). The mRNA levels of group B clades seem to be influenced by a minor quantity of parameters than the mRNA of group A clades, with NH_4^+ being of specific importance. This may indicate an increase of diatoms ability in optimizing ammonium uptake facing a broader spectrum of environmental conditions. Interestingly, clade 3, owned by araphid pennate diatoms, is influenced solely by iron and NO_2^-. Clade 7, which is basal to group A, is specifically influenced by temperature, being strictly located in Antarctic stations.

Concerning di-*NRT2*, the distribution of clade-related environmental parameters for optimal nitrogen uptake is less clear (Fig. 5.10). The basal clade 12 shows a very high contribution of NO_2^- concentration to optimal mRNA abundance, which indicates that ancestral di-*NRT2* mRNA level was specifically dependent on the substrate. Response curves indicate the environmental condition enhancing mRNA abundances and they detect, for the same clade 12, low-medium iron and intermediate nitrates availabilities as optimal conditions (Fig. 5.12). The optimal mRNA abundance of clade 1 and clade 3 di-*NRT2*, the latter including exclusively raphid pennate diatoms, is also based on a restricted number of parameters, indicative of some form of specialization. Overall, also in the case of di-*NRT2*, rounds of gene duplication probably enhanced the ability of diatoms to uptake nutrients in a broader range of environmental conditions. Indeed, for both di-*AMT1* and di-*NRT2* the clades

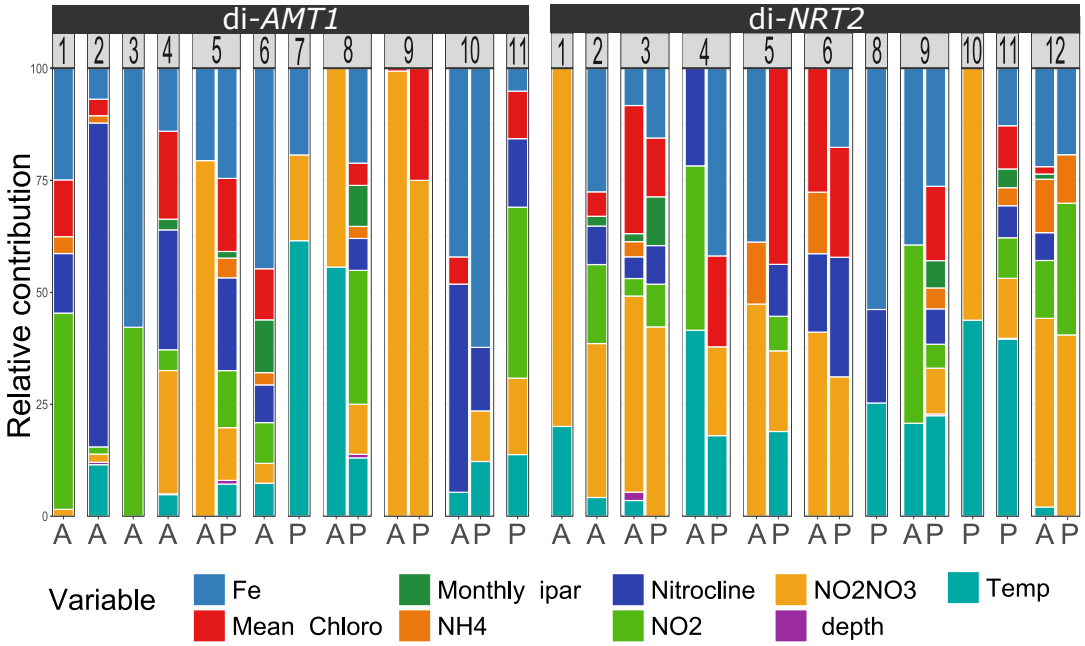

Fig. 5.10: Contribution of environmental predictors in detecting clades optimal conditions, from the BRT models based on both clades presence-absence (P) and mRNA abundance (A).

mostly used in low $NO_2^- + NO_3^-$ availabilities are all located in the most recent clades: di-*AMT1* clades 1, 2, 6 and di-*NRT2* clades 1, 2 and 5.

Temperature

In order to understand how the spatial distribution of N transporters may evolve in the next future in response to climate change, the sensitivity to temperature of the different clades was tested within the BRT modeling. A relatively similar exercise was performed recently by Barton et al. (2010) on the base of diatom taxonomical distribution while Mock et al. (2017) explored the response to temperature by focusing on specific genes. Only the clades presenting temperature among the explicative variables kept by the simplifying process could be included in this exercise. The other are thus to be considered as non responding to the temperature footprint of a climate change. Taking into account climate change, we know that applying different emission scenarios to the CMIP5 model (RCP2.6, RCP4.5, RCP8.5)

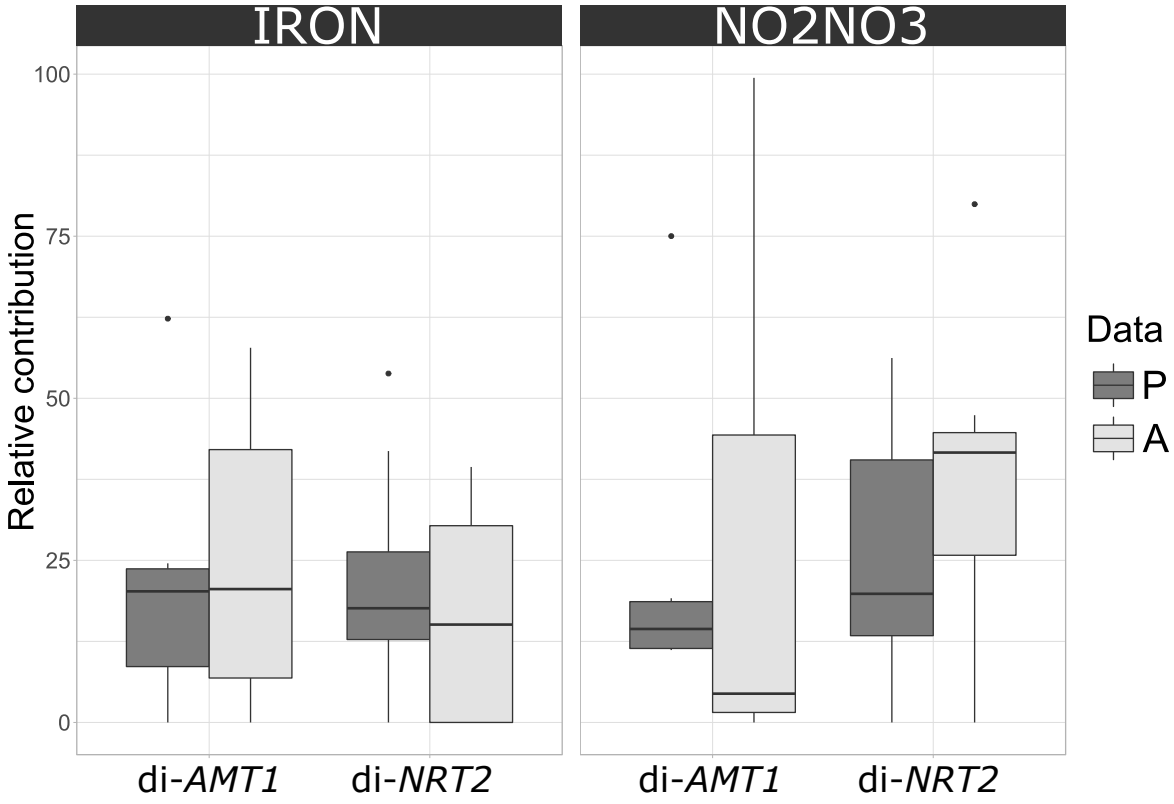

Fig. 5.11: Boxplot of the relative contribution of the two most contributing environmental predictors: iron and $NO_2^- + NO_3^-$ availability for the models based on presence-absence (P) and mRNA abundance data on the 20-180 μm size fraction (A).

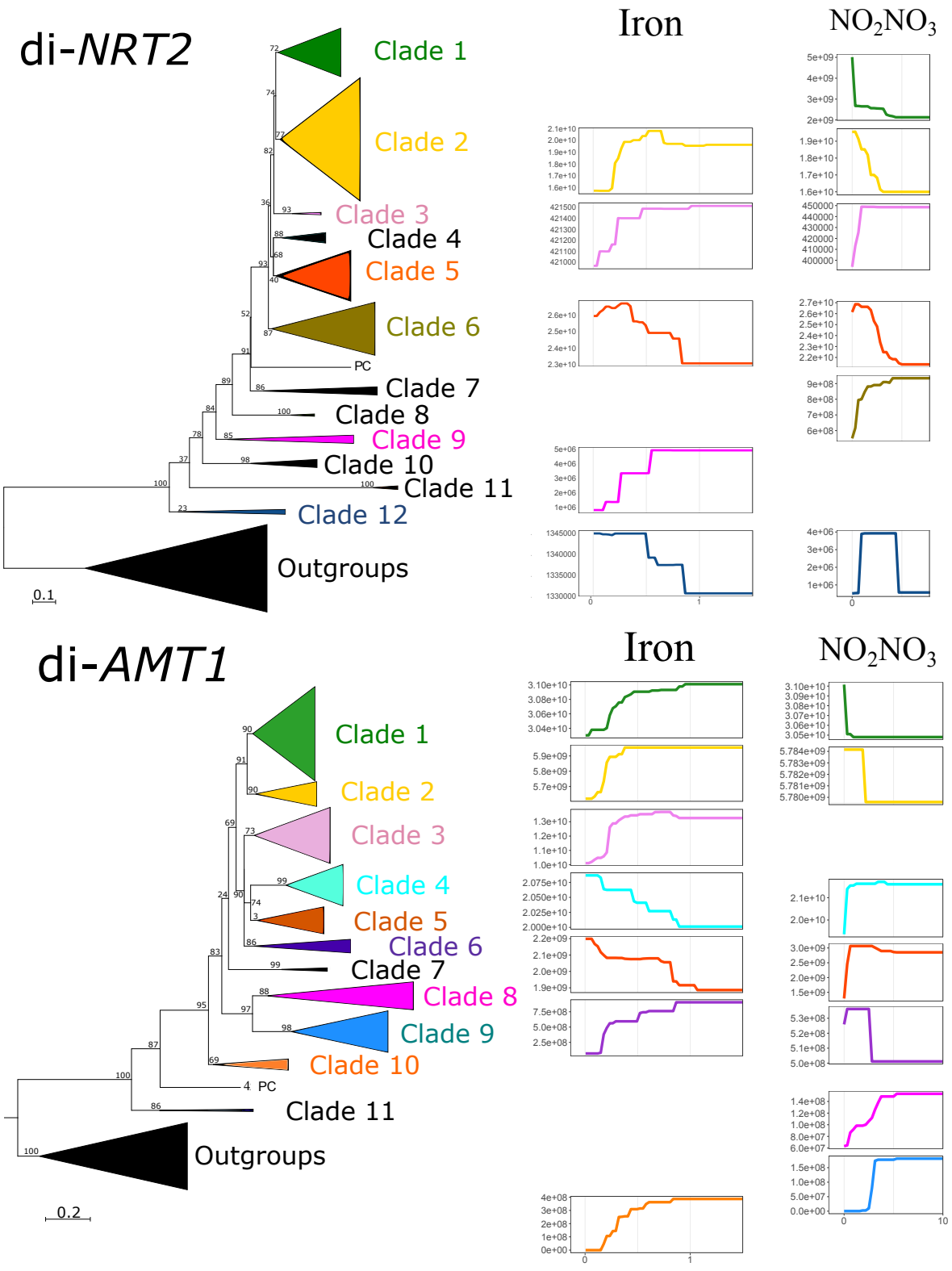

Fig. 5.12: Response curves derived from the BRT models based on mRNA abundances for the two most contributing variables of the models: iron and nitrates. The curves have the same colors of the clades they refers to, depicted on the original phylogenetic tree on the left.

we can obtain global average anomalies respectively of +0.94°C, +1.75°C and+3.75°C in 2086–2100 compared to 1998-2012 (Power et al., 2017). Following these projections I run a simple sensitivity analysis to investigate the clades sensitivity to temperature. Keeping all the other variables fixed to the observed values I predict the probability of presence of each clade in every sampled station with an increase of temperature up to 3.0°C, every 0.5°C steps (Fig. 5.13). Concerning the di-*AMT1* family, increased temperature will have a clear deleterious effect on the distribution of clade 7. This clade, which includes only radial-centric-basal-Coscinodiscophyceae, is basal to group A di-*AMT1* and is distributed in Antarctic regions, thus this result is expected. The response of di-*NRT2* to global warming is more complex. Di-*NRT2* clades were indeed the most temperature-sensitive ones. Three clades (clades 9, 10 and 11) show strong negative response to increased temperature. As these three clades are all basal to supergroup A di-*NRT2*, one may speculate that early di-*NRT2* were not preferentially adapted to high temperature. This analysis does not claim to realistically predict future scenarios as only the temperature variable was taken into account. Importantly, it highlights also that high quality future predictions of other key parameters, such as iron or nitrates, are required to have reliable predictions on the response of diatom functionality.

5.5 Conclusions

While the results have been discussed in the previous text, I add here some general considerations. N transporter clades are not only characterized by different distributions (chapter 4) but also by different modulations as answer to several environmental variables. Through the framework of analysis here developed I was hence able to describe the optimal conditions for the use of each clade (Tab. 5.8), and thus indirectly prove the functional difference of the designed clades.

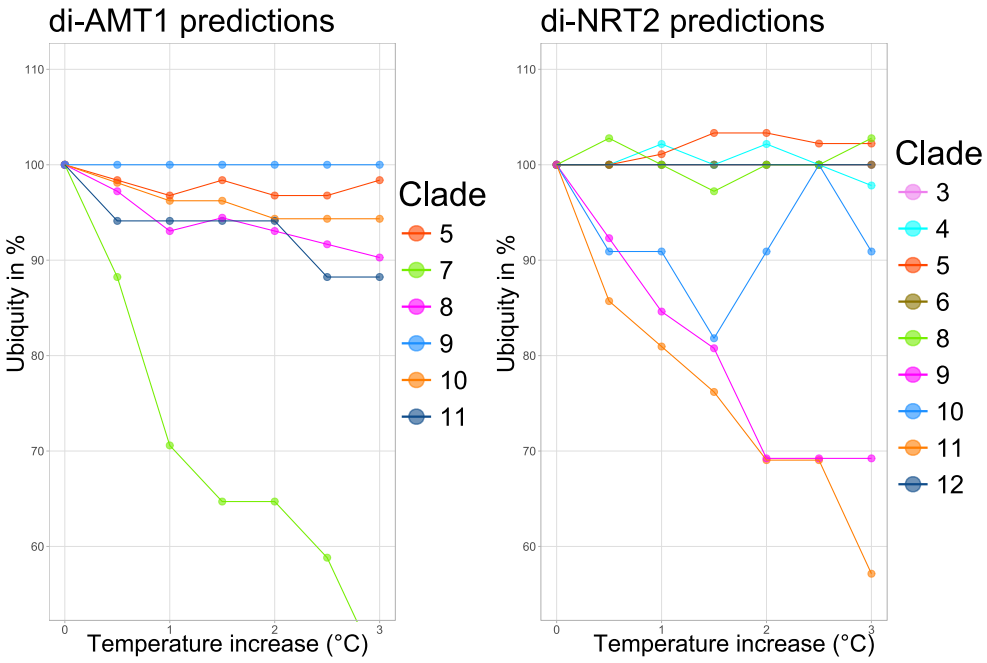

Fig. 5.13: Sensitivity exercise on clades BRT models. Prediction of ubiquity changes on the *Tara* Oceans sampling stations varying only the temperature parameter up to 3.0°C every 0.5°C. Ubiquity is expressed in percentage as the change of ubiquity in the scenario compared to the observed ubiquity, normalized over the observed data itself.

The description of all the clades is summarized in Table 5.8 but I'll report here examples of few key clades emerged from the analysis. Clade di-*AMT1* 10 probably retains the most basal characteristics within the inferred di-*AMT1* phylogeny. Genes belonging to this clade are found only in radial-centric-basal-Coscinodiscophyceae that are preferentially inhabiting the Mediterranean Sea. The distribution and modulation of this clade is strongly influenced by iron availability, for which they need medium-high concentrations. On the other hand, the generalistic clade di-*AMT1* 3, which is owned by organisms from all the major diatom groups, is ubiquitously distributed across the oceans although it is only dominant in the Indian Ocean and within small-medium diatoms (0.8-20 μm). The modulation of genes included in this clade is clearly defined by two nutrients, namely iron and NO_2^-. The basal di-*NRT2* clade 7 is strongly influenced in its mRNA levels by NO_2^-. N related variables are the major modulators also in the case of clade di-*NRT2*-1 and di-*NRT2*-5 which are mostly used in limited N availability, while for other clades temperature, iron and chlorophyll α also play a very important role.

This global survey is strongly suggestive of complex evolutionary scenarios for diatoms N transporters. The lack of taxonomic resolution in our dataset prevented me from deeply analyze the adaptation strategies adopted by distinct species. Overall, diatoms locally deploy a more extensive repertoire of genetic solution (clade richness) for ammonium uptake compared to nitrate one. This is also reflected by the fact that the mRNA abundances of three over twelve di-*NRT2* clades are worldwide dominant, while di-*AMT1* transcripts abundances are widely distributed across clades and regions. This probably indicates that diatoms *AMT1* are more specialized to local conditions. As expected (Marañón et al., 2013), size-class effect was detected in the differential use of N transporters between small and medium/large diatoms with the latter showing a preferential use of a partially different set of genes.

In this study I explored the potentiality of marine meta-omics for the in depth analysis of a functional trait in diatoms. My working hypothesis was that evolutionary differentiations in N transporters was mirrored by patterns of specific adaptations to environmental variables. Although the lack of resolution of our starting dataset did not allow me to perform a fine-scale analysis of gene modulation, results seem to partially support this hypothesis. The use of clades richness as functional measure gave as result reasonable patterns and the characterization of each clade resulted in specific different modulation of mRNA levels. The differential use of environmental cues by different clades supports the sub-functionalization of the latter and, consequently, a differential functional role of each unit, supportive of the working hypothesis. The use of *AMT1* or *NRT2* produced equivalent patterns in clade richness, indicating the interchangeable value of the two high affinity N transporters as functional markers through this framework. Future efforts should focus on the application of the same framework to other key gene families, as for example nitrate reductase, involved in N assimilation.

Tab. 5.8: Resuming table of the information achieved on di-*AMT1* clades through phylogenetic, biogeography and modulation analysis across chapters 3, 4 and 5. For each clade it is described the taxonomic assignation (n° of genes), the ubiquity (n° of stations), the environmental drivers of presence-absence distribution, and the preferential expression use (by which size-class of diatom, where) and modulation (the environmental drivers of expression.

			DISTRIBUTION		EXPRESSION			
	Clade	**Taxonomy**	**Distribution (ubiquity)**	**Environmental drivers (conditions enhancing presence)**	**Size**	**Where**	**Biogeography**	**Environmental drivers (conditions enhancing expression)**
di-*AMT1*	**1**	RC (6), PC (78)	Ubiquitous (101)	/	Big diatoms (20-180 µm)	Across all the basins	di-*AMT1* red	NO_2^-(low availability)
	2	RC (18), PC (4)	Ubiquitous (101)	/	Medium small diatoms (0.8-20 µm)	Expressed across all the basins and dominant in MS, NAO and NPO.	di-*AMT1* yellow	Nitrocline depth (lower depths)
	3	RC(5), PC(5), RP (36), AP (4)	Ubiquitous (104)	/	Small-medium diatoms (0.8-20 µm)	Southern oceans (SO, SPO and SAO) and dominant in the IO.	di-*AMT1* blue	Iron (medium availability) and NO_2^-(low availability)
	4	RC (27), PC(10)	Ubiquitous (92)	/	Medium-big diatoms (5-180 µm),	Expressed in all the basins except the MS, dominant in SAO and SPO.	di-*AMT1* cyan	/
	5	RC (3), PC (21)	Ubiquitous except MS very oligotrophic stations (77)	/	/	Relevant abundances in all the basins except MS and IO.	di-*AMT1* cyan	NO_2^-+NO_3^- (medium availabilities)
	6	RC (6), AP (6)	Ubiquitous except North IO and W-MS (77)	/	/	Specifically expressed in SO and southern IO.	di-*AMT1* pink	Iron (high abundances)
	7	RC (3)	Mainly in SO (8)	Temperature (<15°C)	/	Specifically expressed in SO.	di-*AMT1* pink	/
	8	RC (5), PC (13), RP (7)	Absent only in IO and central Atlantic O (78)	NO_2^-+NO_3^- (medium-high availability) and temperature (<15°C)	Small diatoms (0.8-5 µm)	Never particularly abundant, found in the southern regions (SO, SPO, SAO)	/	/
	9	RC (3), PC (21), RP (13)	Found on nutrient enriched stations (close to upwelling or SO) (48)	NO_2^-+NO_3^- (medium-high availability)	Small diatoms (0.8-5 µm)	Abundant in the SO, for 0.8-5 µm also in other southern regions (SPO and SAO)	di-*AMT1* pink	NO_2^-+NO_3^- (high availability)
	10	RC (10)	MS, East-NAO, coastal areas off Argentina and SouthAfrica (45)	Iron (high availability)	/	Specific of the MS.	di-*AMT1* blue	Iron (high availability)
	11	PC (2)	Only 4 stations in SAO and NAO. (7)	NO_2^-(high availability)	Small diatoms (0.8-5 µm)	SAO and NAO.	/	/

Tab. 5.9: Resuming table of the information achieved on di-*NRT2* clades through phylogenetic, biogeography and modulation analysis across chapters 3, 4 and 5. For each clade it is described the taxonomic assignation (n° of genes), the ubiquity (n° of stations), the environmental drivers of presence-absence distribution, and the preferential expression use (by which size-class of diatom, where) and modulation (the environmental drivers of expression.

			DISTRIBUTION		EXPRESSION			
	Clade	**Taxonomy**	**Distribution (ubiquity)**	**Environmental drivers (conditions enhancing presence)**	**Size**	**Where**	**Biogeography**	**Environmental drivers (conditions enhancing expression)**
di-*NRT2*	**1**	PC (4), RP (36), AP (2)	Ubiquitous (92)	/	Small-medium diatoms (0.8-20 µm)	Relevant everywhere except SO, dominant in IO, MS and NPO.	di-*NRT2* blue	NO_2^-+NO_3^- (limited concentrations)
	2	RC (9), PC (91), RP (6), AP (1)	Ubiquitous (100)	/	All size fractions.	Very abundant across all the basins (except SO)	di-*NRT2* red and di-*NRT2* cyan	/
	3	RP (2)	NAO and off the S-Africa both in the IO and in the SAO (25)	NO_2^-+NO_3^- (low concentrations)	Medium-big diatoms (5-180 µm)	Rarely abundant only in the IO	/	NO_2^-+NO_3^- (medium-high concentrations) and iron (medium-high availability)
	4	PC (5), AP (4)	Spread across the basins except the Pacific Ocean (50)	Iron (medium-high availability)	Medium diatoms (5-20 µm)	NAO.	/	Temperature (<17°C) and nitrocline depth (depths >180m)
	5	PC (23), RP (3), AP (18)	Ubiquitous (96)	/	Medium-big diatoms (5-180 µm)	Across all the oceans except SO.	di-*NRT2* yellow, di-*NRT2* cyan	NO_2^-+NO_3^- (low availabilities)
	6	RC (4), PC (16), RP (19), AP (7)	Ubiquitous except very oligotrophic stations and MS (79)	/	/	Dominant in SO, enriched in several stations of SAO and SPO.	di-*NRT2* pink	NO_2^-+NO_3^- (medium-high concentrations)
	7	PC (2), RP (1), AP (4)	Spread in stations across all the basins (44)	/	/	Abundant in very few stations located across all the basins.	/	/
	8	RP (2)	Specific of SPO but also found in 3 stations spread in NAO, SAO and IO (18)	Iron (low concentrations)	Medium-big diatoms (5-180 µm)	Extremely rare in across NPO,SPO and NAO	/	/
	9	RC (1), PC (6)	Few stations spread across the oceans but mainly S Africa and MS. (24)	/	/	Relevant abundances in the MS, in 92_SRF (SPO) and 66_SRF (SAO)	/	/
	10	PC (4), AP (3)	Only in SO (7)	NO_2^-+NO_3^- (high availability) and temp. (<15°C)	Medium-big diatoms (5-180 µm)	Typical of SO	di-*NRT2* pink	/
	11	PC (1), RP (2)	Only higher latitudes stations, mainly in the Atlantic Ocean (34)	Temperature (15-20°C)	Small diatoms (0.8-5 µm),	Dominant in MS and SPO (only DCM), but present also in NAO, SAO and NPO.	Absent in 20-180 µm	
	12	RC (2), PC (2)	Everywhere except MS and IO (38)	NO_2^-+NO_3^- (medium-high concentration) and NO_2^-(medium availability)		Relevant abundances only in 4_DCM (NAO) and 100_DCM (SPO)	/	NO_2^-(high concentrations)

6

Diatoms: from omics to conceptual models

6.1 Summary and main achievements

- In this chapter I performed a comparison of global phytoplankton numerical modeling to metabarcoding data from *Tara* Oceans;
- I validated the enviromental context of the implemented MIT model;
- I found multiple correspondences between metabarcoding units (OTUs) and numerical modeling units (types);
- I investigated the parameterization of numerical modeling units associated to OTUs in order to obtain physiological information about OTUs;
- Diatoms OTUs are found to have a broader range of physiological characterizations compared to the ones they are usually modeled with, highlighting the incredible diversity associated to this taxa;
- Physiological characterization of OTUs allowed me to investigate size-class differentiations of diatoms and possible different functional strategies in nutrient uptake.

6.2 Introduction

Plankton studies have been deeply affected by the dramatic increase of genomic and metagenomic data availability. Indeed, these approaches arose as extremely promising tools to give access to a finer understanding of the plankton diversity and functioning (chapter 1.1). The question now more than ever is how to proceed with it.

Through this thesis you have read about several efforts in making the most of omic data to investigate diatom diversity. The analysis of omic data by ordination methods, statistical and machine learning approaches allowed insights on diatom ecology from different points of view. A completely different field aiming to provide deep understanding of the processes controlling the plankton system is numerical modeling, that is, the computer-based simulations of diatom populations' changes in time and space. These exercises are based on sets of mathematical equations that are often a combination of the mathematical translation of conceptual ecological and biological models and of an empirical fitting of equations to observed laboratory and *in situ* data (the so-called parameterization, see Follows and Dutkiewicz, 2011 and references therein). To make an example of the latter let's consider the grazing pressure: this process is often expressed as the product of the prey and predator abundances, multiplied by constants expressing prey selectivity, ingestion rates etc. The biological and ecological features of a specific group are thus summarized by a specific set of values for the various constants, constants which are often *ad hoc* adjusted to have reasonable results (model tuning). Within this context, as the diatoms classical view is of ecosystem opportunists (fast growing when resources are abundant), they are generally parameterized with high maximal growth rates and high half saturation constants for nitrate. Furthermore, in numerical models they are likely characterized by the limitation by silica and/or iron availability. At their best, numerical models are very useful to test hypotheses that are hard to develop experimentally since

plankton communities are the emergent properties of many non-linear interactions among the many elements of the complex marine ecosystem (physics, chemistry, biology, etc). More often, they are used for hindcast or prediction in realistic conditions. They are indeed fundamental for our ability to predict the response to greenhouse gases because of the strict linking between nutrient cyclings and phytoplankton (chapter 1.2.1). Nevertheless, the current generation of biogeochemical models does not incorporate, e.g., the plasticity, such as nutrient co-limitation (Browning et al., 2017) and uncertainties in the model formulations, such as the representation of critical bio-limiting resources like iron (Tagliabue et al., 2016). This prevents reliable predictions, especially for fundamental properties like primary production (Frölicher et al., 2016; Bonan and Doney, 2018) and ultimately food sustainability and predictive capabilities (Bonan and Doney, 2018). To overcome current limitations it has been proposed to build on the augmented observations and to use novel taxonomic and functional genomics (Coles et al., 2017; Stec et al., 2017), leading to advanced high-resolution, hydrodynamic-biogeochemical-genomic models.

As you could read in Chapter 1.1.3, several numerical modeling exercises dealt with global-scale planktonic diversity, with a still open debate on how to represent the actual diversity under the constraints of limiting the number of "species" or ecosystem units. At this regard, plankton diversity is nothing more than the result of single species distribution in space and time. This distribution can be interpreted as the realized niche of the species themselves, that finally is the result of abiotic preferences and biotic interactions (e.g., competition, predation). Understanding all the processes structuring the niche is fundamental for ecosystem modeling in order to draw a correct conceptual framework of the marine planktonic ecosystems (Stec et al., 2017). As both omic and modeling approaches have the potentiality to provide insights on the dynamics of these communities, these two approaches could take advantage from each other.

The integration of modeling and omics approaches has hence enormous potentiality but it is actually in its infancy (Coles et al., 2017). While the modeling approach tries to summarize ecosystem dynamics in several equations and parameterizations, the omic is the *in situ* observation of the whole community: a complete thorough picture of the punctual situation. Nevertheless, even if modeling has a conceptual scheme of the dynamics it lacks of a sufficiently detailed description of the units: modeled phytoplanktonic types are characterized by several assumptions or approximations. Growth rate, optimal temperature, nutrient uptake rates and so on are generally extrapolated from the study of a few model (i.e., easy to manage in the laboratory but not necessarily ecologically relevant) species, and often averaged. Because of this, these units do not correspond to any specific taxonomic assignation. On the contrary, the omic approaches allow a taxonomic assignation of the elements present, limited by the references availability (chapters 2 and 3), and it potentially informs on the activity of the units (metatranscriptomic; chapter 5) but it is not possible to understand comprehensive community-level dynamics from this information. Finding a correspondence between the OTUs derived by a metabarcoding dataset and the plankton types implemented in a numerical model would thus allow to infer properties (i.e., parameterizations, such as growth rate) of the OTUs. Such inference may be very delicate to do as it is based on several assumptions: that the model is correctly representing the spectrum of physiologies locally present and, more generally, that the model is correctly simulating the *in situ* conditions. However, if these conditions are met, we would be able to physiologically characterize the OTUs and thus phytoplanktonic taxa, most of which is still unculturable in the laboratory and, consequently, for which direct ecological and physiological characteristics are still impossible to retrieve. The important question now is: can real diatom be associated to (and only to) opportunistic model types? This exercise, never done before, is within the aims of this chapter.

Furthermore, virtually, associating the units of the model to observed OTUs would give insights also on the model parameterizations. A possible limitation toward this direction is that, technically, different combinations of model parameterizations could produce types characterized by similar realized niches. However, keeping this limit into consideration, the combination of omic and modeling data could help answer the following question: are the modeled diatoms similar to real diatoms or do we need to change the conceptual (hence, numerical) view?

The integration of these two kinds of data is thus crucial both for modeling work, to validate their results, as well as for traditional ecologist and physiologist, to assess the distribution and possible physiologies of taxa rarely studied in the wetlab. In this chapter I will search for correspondences between diatoms as described by the *Tara* Oceans metabarcoding and the 350 phytoplanktonic types included in a model developed at the Massachusetts Institute of Technology (MIT), currently by far the global model with the highest diversity of types. The work has to be intended as preliminary given the complexity of the two datasets and the related limitations.

6.3 Material and Methods

6.3.1 Data

The model

The numerical model outputs were provided by Dr. Stephanie Dutkiewicz (MIT), who collaborated to this exercise. The model is strongly based on the model published in Dutkiewicz et al. (2015), all the applied algorithms can be found on this latter, with just a higher number of modeled phytoplankton units.

The biogeochemical equations are the same of the Darwin model (Follows et al., 2007; Dutkiewicz et al., 2009; Hickman et al., 2010; Dutkiewicz et al., 2012). It considers carbon, phosphorus, nitrogen, silica, iron and oxygen cycles through their different phases (i.e., inorganic, living, dissolved, and particulate organic). Atmospheric iron inputs in the ocean surface are from Luo et al. (2008). The ocean circulation model used is the MIT general circulation model (MITgcm) (Marshall et al., 1997) in a global three-dimensional configuration constrained to be consistent with observations (the ECCO-GODAE state estimates; Wunsch and Heimbach, 2007). The resolution has 1 square degree spanning 23 depth levels (from 10 m to 500 m depth).

The model uses a complex marine ecosystem incorporating several phytoplankton and zooplankton types. Totally, the phytoplankton community is described in the model by 350 types, while zooplankton by 16 types. Phytoplankton types can be described within three dimensions of trait space: size, biogeochemical function, and temperature tolerance. Within the 350 phytoplanktonic types, 110 types are characterized as diatoms through their use of silica. Phytoplankton functional types are parameterized to simulate diatoms, large eukaryotes, coccolithophores, mixotrophic dinoflagellates, picoeukaryotes, *Synechococcus*, high- and low- light *Prochlorococcus*, nitrogen-fixing *Trichodesmium*, and unicellular diazotrophs. The main differences among these phytoplankton are parameterized through different elemental requirements: maximum growth rate, nutrient half-saturation constants, sinking rates, maximum Chl α:C and palatability to grazers. Parameter values are distributed stochastically in the model but the size classes assigned to phytoplankton types define some trade-offs amongst them as several processes are assumed to be governed by size (e.g., smaller cells have lower sinking speed and nutrient half-saturation constants). Moreover, functional groups (following Dutkiewicz et al., 2015) have specific differences for maximum growth rates, cell elemental stoichiometry, and palatability to grazers.

Biogeochemical and biological tracers interact in terms of organic matter flux: i.e., according to formation, transformation and remineralization processes of the organic matter. Consequent to processes like excretion and mortality, living organic material is transferred into sinking particulate and dissolved organic detritus, which are transformed back to inorganic material through respiration processes. Phytoplankton performances are determined by their interaction with the environment. Consequently the model shapes the ecosystem structure and the feedback of these interactions on the resources. There is thus a self-organisation of the communities, with the simulated patterns emerging from a large set of possibilities.

The abundances of all the phytoplankton types have been extracted from the model output in terms of biomass (mmolC/m^3) and are submitted as Supplementary File 14, and the traits characterization of each type is presented in Supplementary File 15.

6.3.2 The model output mining

Only surface values were considered, that is at the first depth level (0-10 m) of the physical model, for all the *Tara* Oceans sampling site locations and for the month corresponding to the sampling time of the expedition. Bray-Curtis distances between phytoplankton types were run through the R *vegan* package (Oksanen et al., 2017) and plotted in a heatmap (package *pheatmap*, Kolde, 2015).

To compute the richness of types over the *Tara* Oceans samples a threshold of presence was assessed a priori, equal to $1e^{-23}$, 20 orders of magnitude higher than the lowest allowed modeled abundance. The richness distribution across the latitude was compared to the diatom richness derived in chapter 2 based on the filtered and the unfiltered datasets.

6.3.3 The metabarcoding

The metabarcoding dataset of *Tara* Oceans has been already described in chapter 2. For this exercise I will exploit the Swarm Metabarcode, clustered with a clustering level equal to 1 and filtered at threshold equal to 99.65 (see chapter 2). OTUs abundances are normalized over the total abundance of OTUs measured in the sampling.
OTUs ubiquity has been computed as the number of samples where the OTU has been detected as present. OTU median abundances correspond to the median of the normalized abundances where the same is higher than zero.

6.3.4 The environmental comparison

In order to validate the model environmental description (that is, its closeness to observed environmental properties), pairwise Pearson correlation were run between the numerical environmental variables and the corresponding variables as measured *in situ* during the *Tara* Oceans expedition.

6.3.5 The comparison types-OTUs

To establish a correspondence between OTUs and numerical types, pairwise Pearson correlations have been computed between all the OTUs and all the phytoplanktonic types over the *Tara* Oceans sampling system. In order to have comparable data, the abundance of phytoplankton types has been normalized over the total concentration of types in each sample beforehand. Every pairwise correlation has been calculated over the sampling points where both single OTU and type were considered pairwise present and only where the number of sampling where both units were present was of at least of 10 stations. *P*-value was adjusted according to the Benjamini & Hochberg method (1995), and correlations were considered significant only if the ad-

justed p-value was lower than 0.05 and the correlation ρ was higher than 0.4. Significant correlations were then mapped in heatmaps through the R package *pheatmap* (Kolde, 2015) and aggregated by the taxonomic annotation of OTUs in genus. The resulting heatmap shows rows and columns ordered according to a ward.D clustering based on the Jaccard distances of the same. On the same heatmap I mapped the values of two parameters used to characterize the phytoplanktonic types in the model: PCMAX, the maximum photosynthetic rate (roughly, the maximum growth rate), and KSATNO3, the half saturation for growth for nitrate. The maximum growth rate is strictly dependent on the assimilation ability of the organisms, and this parameter alone is informative of the degree of resource utilization (Edwards et al., 2012). Moreover, once again, we focus on the utilization of nitrogen by diatoms since, as previously discussed in chapter 3, this may be a key descriptor of the resource utilization trait. This last constant, K for nitrate, corresponds to the concentration supporting an uptake rate one-half the maximum rate, and it is a measure of the species ability to use low availability of this nutrient (Eppley et al., 1969). This constant varies according to the cell size and to the specific growth rate, and it is inversely associated to the nitrate transporters efficiency (Edwards et al., 2012; Beltrán-Heredia et al., 2017). Furthermore, considering the same number of transporters and of external nitrate concentration, the assimilation rate depends on transporters efficiency (handling time) and on its ability to optimize the meeting with the substrate (Aksnes and Egge, 1991; Aksnes and Cao, 2011).

After establishing which type correlates with which OTU, values of PCMAX and KSATNO3 were assigned to each OTU per size class, calculated as the median of the corresponding values of the model phytoplankton types found to significantly correlate to the same OTUs in the same size class. The resulting 'parameterization' of OTUs PCMAX and KSATNO3 has been compared among the different size classes through a pairwise Student t-test. To test if more ubiquitous diatoms are less prone to depend upon high nutrient values, OTUs

KSATNO3 were then compared to OTU ubiquity. Furthermore, after partitioning KSATNO3 values into 15 different classes ($x>3$, $2\leq x<3$, $1\leq x<2$, $0.9\leq x<1$, $0.8\leq x<0.9$, $0.7\leq x<0.8$, $0.6\leq x<0.7$, $0.5\leq x<0.6$, $0.4\leq x<0.5$, $0.3\leq x<0.4$, $0.2\leq x<0.3$, $0.15\leq x<0.2$, $0.1\leq x<0.15$, $0.05\leq x<0.1$, $x<0.05$), a KSATNO3 functional richness was estimated computing the number of different classes of KSATNO3 assigned to the OTUs present in each sample.

6.4 Results and Discussion

6.4.1 Model validation

To evaluate the comparability of the *in situ* data and the model data we firstly compared the environmental conditions of both. Notably, if the model simulates an environmental context completely different from those observed *in situ*, it would make no sense to compare abundances from the two datasets. The first comparison is thus on a set of nutrient availabilities (nitrate, nitrite, phosphate and silica), and the two major physical descriptors of the ocean water masses: temperature and salinity (Fig. 6.1).

Temperature and salinity from the model are strongly coherent with what has been observed *in situ* with Pearson correlation ρ higher than 0.9. For what concerns nitrate sources, nitrate is well estimated by the model whereas the nitrite concentration shows no correlation between the modeled and observed values. This is easily explained by the punctuality and rapidity of the processes leading to NO_2^- production. Indeed, within the nitrogen cycle the production of NO_2^- is ruled by the nitrification processes, which recycle ammonium to NO_2^- in case of ammonium availability (chapter 1.2.1).

Other fundamental nutrients for diatoms, like phosphate and silica, also showed very good correspondences between *in situ* and modeled data.

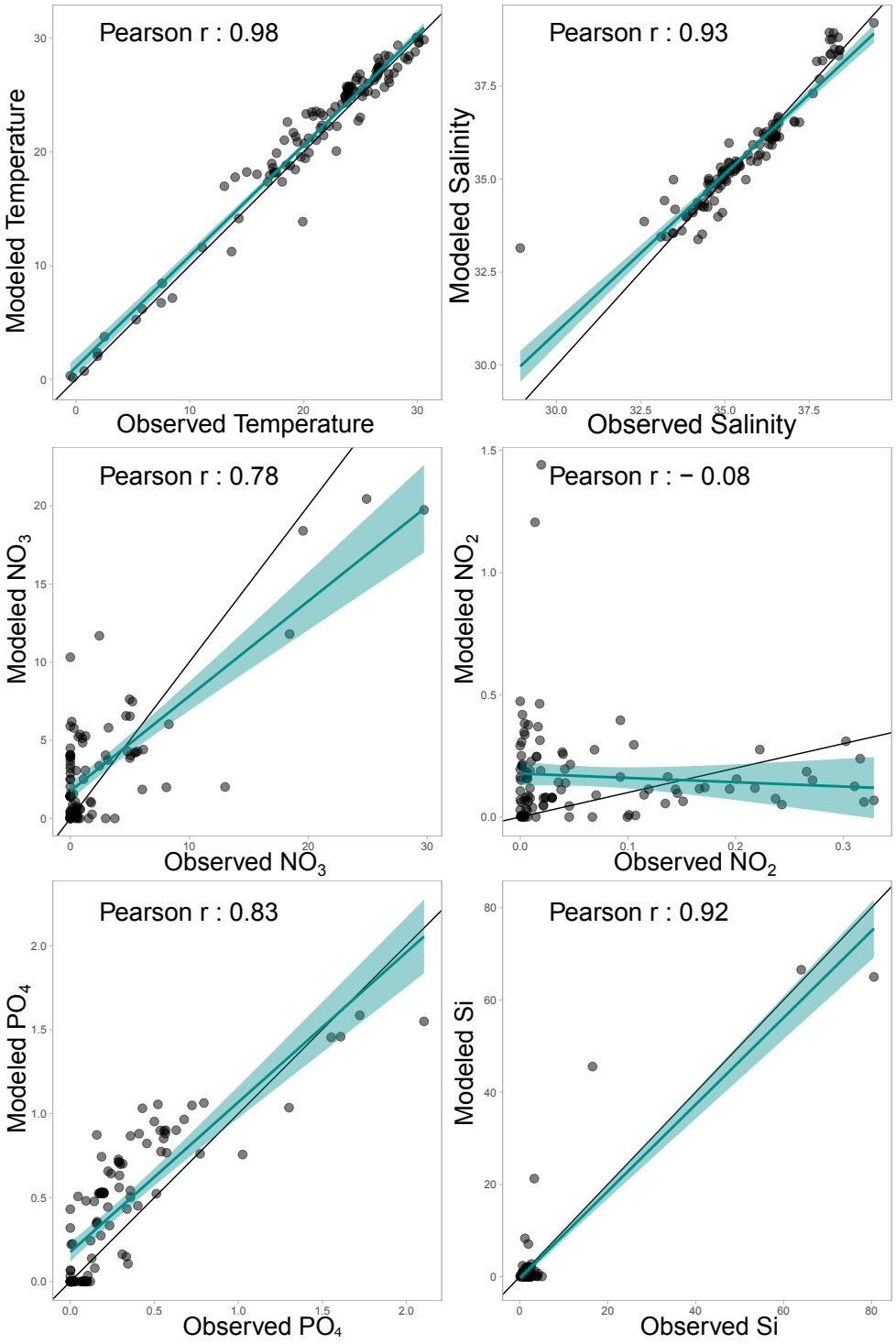

Fig. 6.1: Pearson correlation between the environmental variables as derived from the model and the values of the same variables measured *in situ* during the *Tara* Oceans expedition.

Overall, the model well represents the environmental states observed during the *Tara* Oceans expedition and for this reason I could proceed with the comparison of the phytoplankton units of the models to the taxonomic units of the metabarcode.

6.4.2 Phytoplankton: correspondences between the model and the reality

Bray-Curtis distances were applied to the model phytoplankton types abundances. In this way, I could assess the degree of distinctiveness of phytoplankton types within the *Tara* Oceans sampling context (Fig. 6.2). Notwithstanding the high number of types included in the model, their distances show highly differentiated distributions. Types clustered together into four major groups and, interestingly, the types associated to diatoms do not cluster together but they have very spread kinds of distributions across the distributions of all the phytoplanktonic types. This could suggest the presence of four main distinct physiologies across modeled types characterized by slightly different parameterized types.

To get a picture of the distribution of phytoplankton types over the *Tara* Oceans samples the richness index of the model was measured (Fig. 6.3). Through this thesis I explored the functional (chapter 4) and taxonomic richness (chapter 2) through several approaches. The diatom richness derived by the model presents very strong latitudinal patterns with peaks of diversity at the poles. According to the model, the richness measured on the whole phytoplankton or only within the diatom group show very similar patterns. These strong latitudinal gradients are very similar to the ones observed in the unfiltered metabarcode (Fig. 6.4), with peaks at the poles and a third peak at the equator while at mid-latitudes we find the lower richness. Notwithstanding the similar latitudinal gradient between the two there are strong differences in

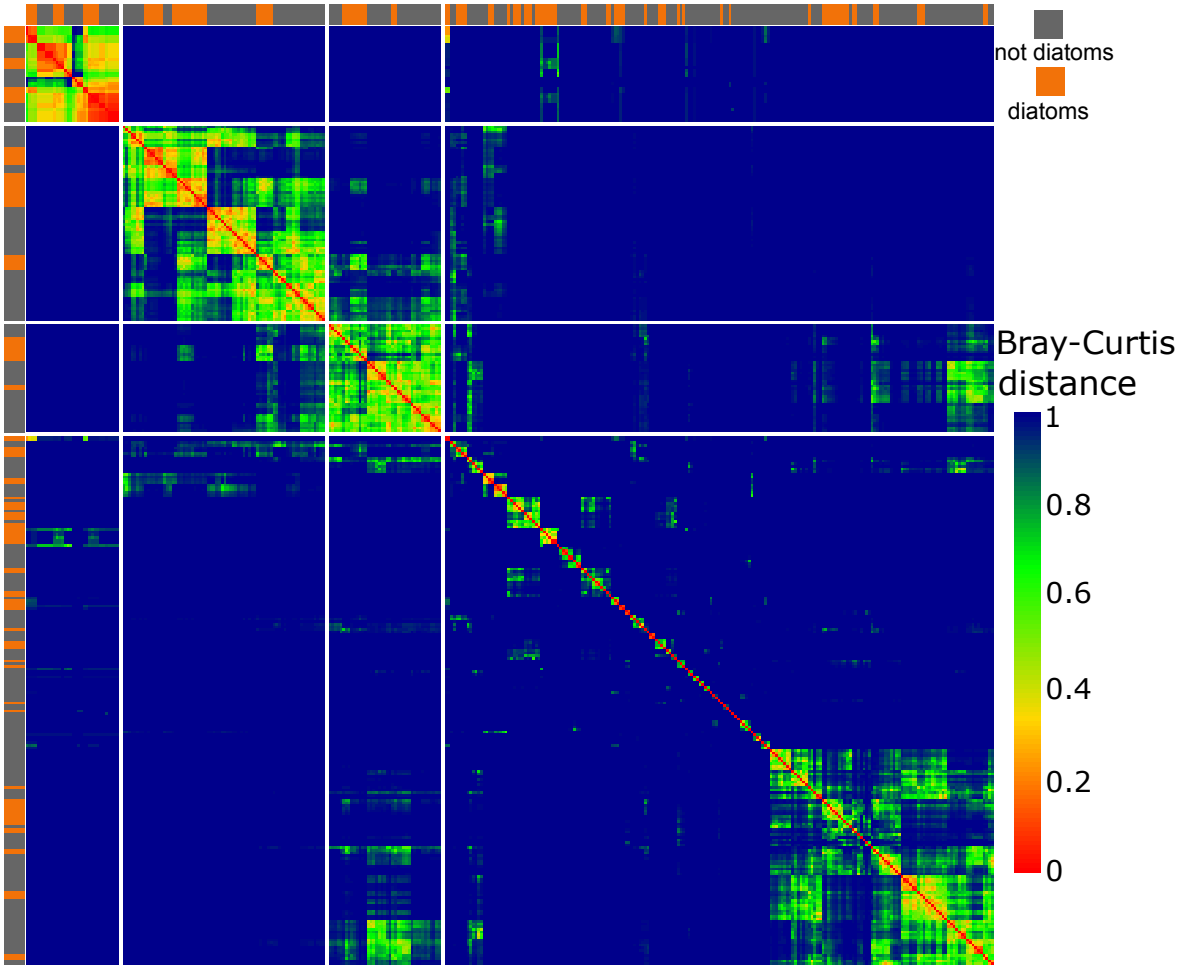

Fig. 6.2: Heatmap of Bray-Curtis distances between phytoplankton types abundances at the *Tara* Oceans sampling sites. Every row and column correspond to a phytoplankton type. The heatmap is annotated according to the type identification: they are colored in orange if they simulate diatoms or in gray otherwise.

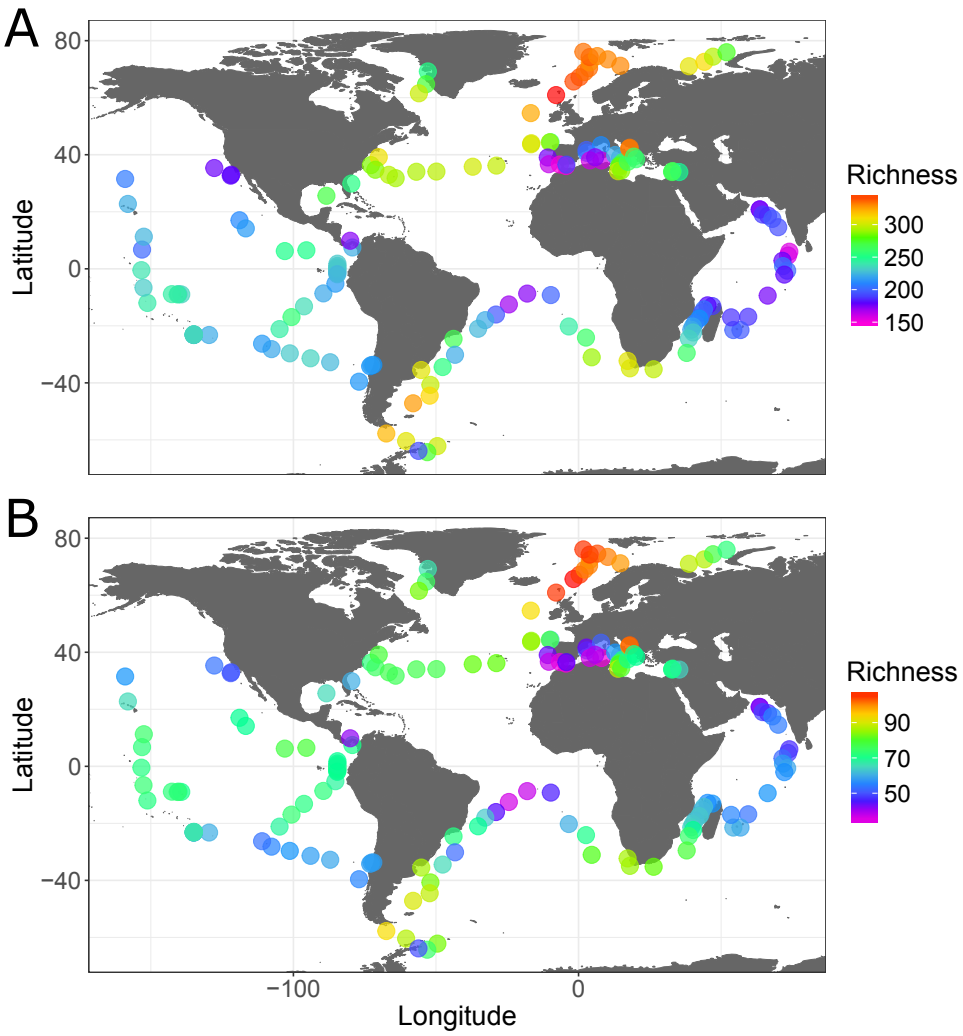

Fig. 6.3: Surface phytoplanktonic types richness based on all the modeled types (panel A) or confined to diatom-simulating types (panel B) across the *Tara* Oceans stations.

longitudinal terms. Indeed, while the model predicted higher richness in the equatorial Pacific Ocean, the metabarcoding data detected higher indices also in the Indian Ocean and in the South Atlantic Ocean. Comparing the model richness to the putative functional richness estimated over the N transporter gene families (chapter 4) gave some noteworthy similarities. Again we found the same peak at the tropics. The comparison at the poles is harder however. Indeed, at the poles we have coherence according to the *AMT1*-based functional richness, but this peak is not confirmed by the *NRT2*-based measure. Moreover, unfortunately in this putative functional richness exercise (chapter 4) the Arctic stations were not included and thus we cannot compare the Arctic peak found on the modeled richness with this other measures.

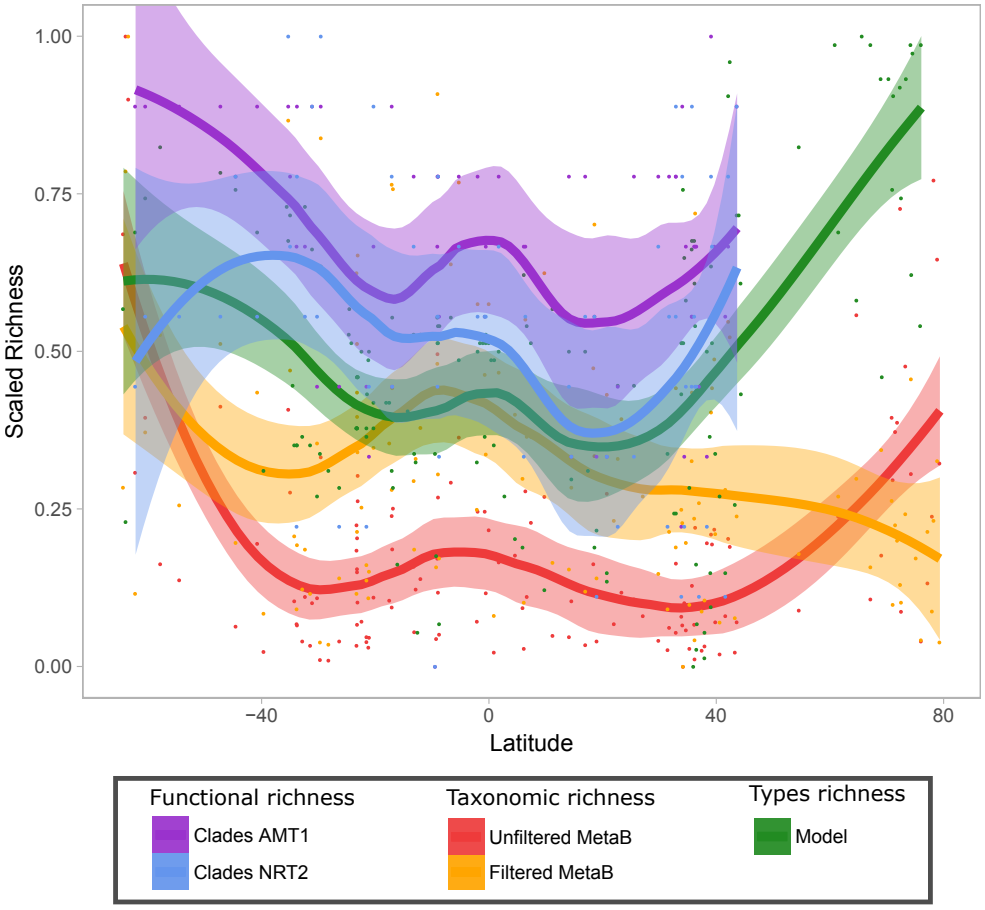

Fig. 6.4: Distribution of diatom richness measures across the latitude. Five different indices of diatom richness have been included: the richness derived from the model (considering only diatom-simulating types), the taxonomic richness obtained by the unfiltered and filtered metaB (chapter 2), and the putative functional richness as based on the *AMT1* and *NRT2* gene families (chapter 4). The richness measures from each dataset have been scaled to 0-1. Across the distributions loess curves are fitted.

Pairwise correlations between phytoplankton types and diatom OTUs within the *Tara* Oceans framework resulted in 6,064 significant correlations, finding at least one significant correlation for 349 phytoplanktonic types over the 350 modeled. Only 180 OTUs found at least one significant correlation over the 402 OTUs that were present in at least 10 samples. These 402 OTUs were annotated to 58 different diatom genus and the subset of OTUs with significant correlation covered 41 of these diatom genus. Overall, the OTUs correlating to the phytoplanktonic types covered most of diatom taxonomic diversity found in the *Tara* Oceans metabarcoding. However, not all the diatoms OTUs extracted have a taxonomic annotation up to the genus level. A number of 1,014 correlations, i.e., around 1/6 of all the significant correlations, is found to be with diatoms OTUs of unknown genus. Looking at the other significant correlations distributed among diatom genus (Fig. 6.5) I obtain a complex set of correspondences. It is expected to see significant correlations with several diatom genus of the same phytoplanktonic type because, by definition, each type is an aggregation of species sharing similar abundances and thus, supposedly, similar physiologies and relations with grazers.

What it is striking it is to see the opposite: to each genus and moreover to each OTU correspond several types. This could be explained at least partially by the different ubiquity of the units compared: while the types have modeled abundances over all the dataset the OTUs are generally rare and the number of shared sampling points over which the correlations are computed is just a subset of the whole dataset. These observation are hence easily explainable by a set of local correspondences between OTUs and the same types, covering different regions each time. In addition, the existence of a significant diatom functional redundancy cannot be excluded. This redundancy could help indeed explain their high diversity (Rosenfeld, 2016).

Furthermore, it is clear that not only the phytoplankton types simulating diatom physiology found significant correlations with diatom OTUs, but also

all the other types exhibited several significant correlations. This finding would question if the physiological characterization of diatom used up to now in global models is actually correct, or if it is only a subset of a wider range of their actual capabilities. The genera finding correspondences to a higher number of types are *Chaetoceros* and *Thalassiosira*, which is expected considering they own a very high diversity within (Malviya et al., 2016). The heatmap shows some consistency in the size-fraction location of significant correlations within the same genus, highlighting the preferential size fractions of each genus. Moreover, sets of very different size fractions are found only on wide genus, covering a high number of very size-differentiated species (e.g., *Pseudo-nitzschia* and *Chaetoceros*).

Strongly, the number of significant correlation is highly dependent on the ubiquity of OTUs. Rare OTUs will be hardly found significantly correlated and the quality filtering itself exclude all the OTUs present in less than five sampling stations. But are the correlations we observe biased by OTUs ubiquity? The distribution of the number of correlations of every OTUs along their ubiquity range (Fig. 6.6) seems to delineate no bias of the correlations over the number of points available to compute the analysis itself.

6.4.3 Diatoms functional diversity

To follow, I investigated the possibility to associate a functional characterization to the OTUs. To do so I selected two major parameters characterizing diatoms: the growth rate expressed by the proxy of PCMAX, the maximum photosynthetic rate, and KSATNO3: the half saturation for growth for nitrate, a fundamental nutrient for this taxa. Going back to the heatmap designed by the presence of significant correlations between OTUs genus and phytoplanktonic types of Fig. 6.5 I focus here on the dendrogram based on the Jaccard distances between the types which ordered the already mentioned heatmap rows. Types which correlate to similar diatom genus OTUs are associated to similar

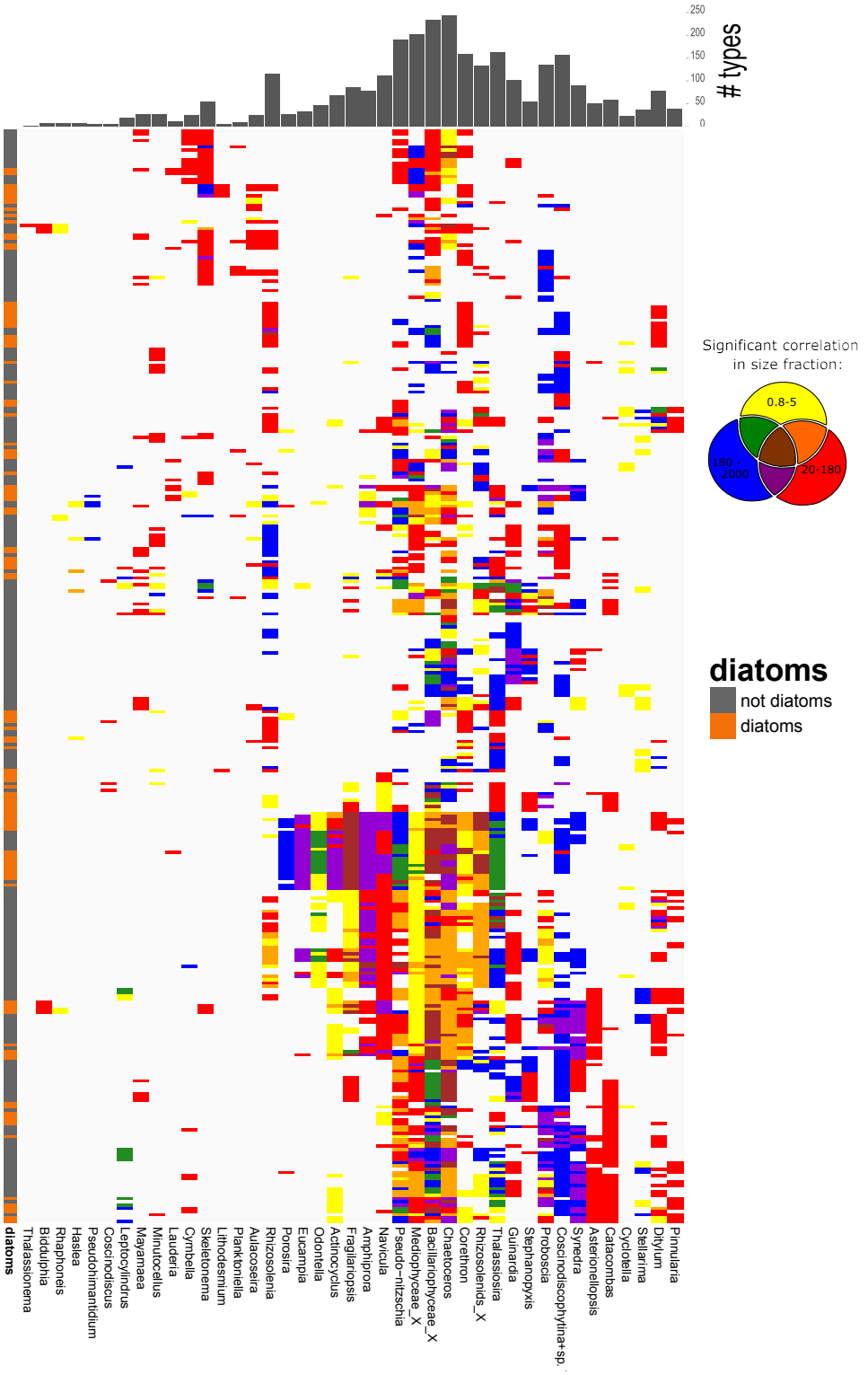

Fig. 6.5: Heatmap of significant correlations found between metabarcoding OTUs and phytoplanktonic types. The x axis corresponds to all the genus found among the significant correlated OTUs, while the y axis corresponds to all the modeled types. On the left, each row, and thus each phytoplankton type, is annotated according to their types characteristics (diatoms or non-diatom). Cells are colored if there is at least one OTU annotated to the corresponding genus to be significantly correlated in at least one size fraction to the phytoplanktonic type in the y axis. The color code corresponds to the size fraction where the correlation was found to be significant, according to the color schema in the right. On the top of the graph a barplot indicates the number of different types found to be significantly correlated with at least one OTUs of the corresponding genus.

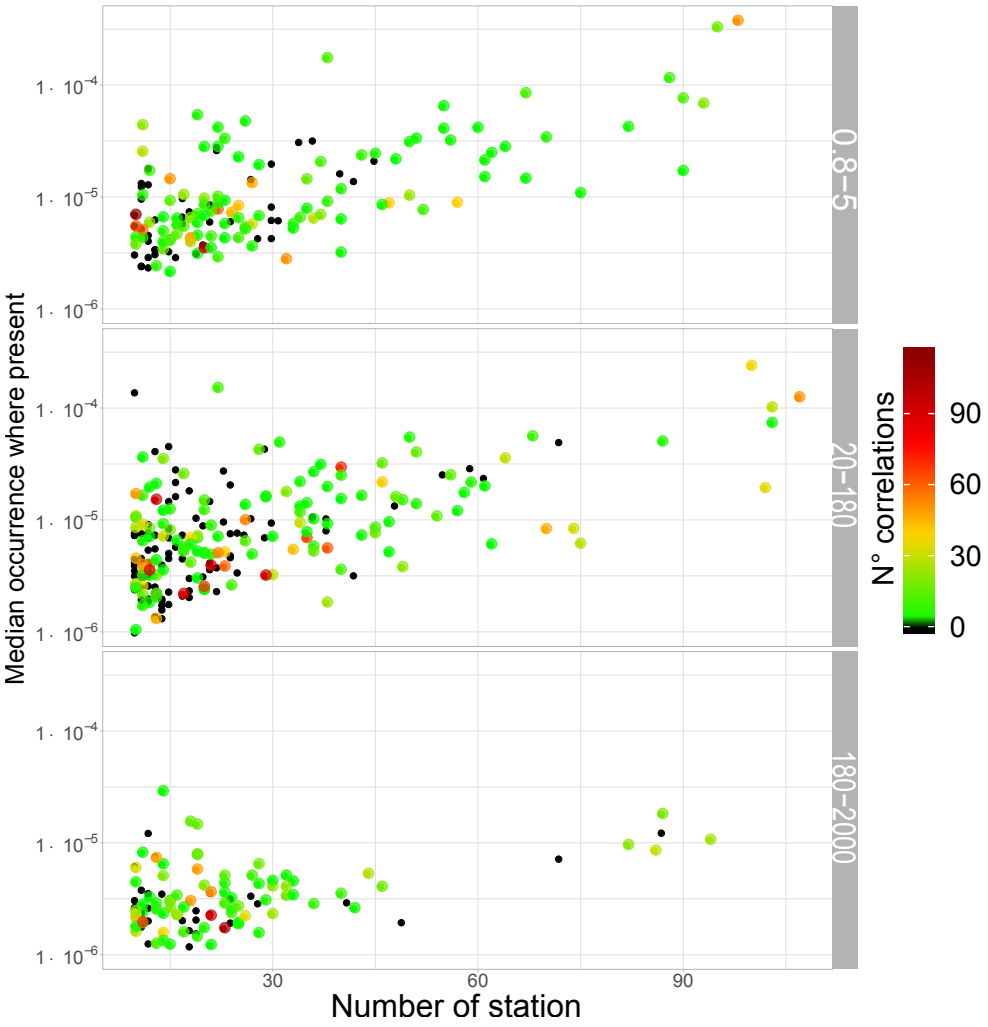

Fig. 6.6: Number of significant correlation per OTUs according to their ubiquity (x axis) and the median abundance where they are present (y axis). Black dots designate OTUs without any significant correlation with phytoplankton types.

parameterization? Comparing the previous heatmap (Fig. 6.6, reported in Fig. 6.7) to the parametrization of types (Fig. 6.7) shows that, generally, types are not grouped according to their growth rate but they are rather clustered by their nitrate half saturation. Again, diatom types are mostly scattered across all the types and they don't have similar PCMAX or KSATNO3 to the other closely clustered types. The lack of clustering on parameterization is explained by the importance of other parameters and by the use of a thermal optimal windowing in the model. That is, two species can have similar growth rate but be forced to live in different temperature ranges.

However, interestingly diatoms simulating types sharing similar PCMAX or KSATNO3 are clustered close to each other, highlighting the importance of these two parameters in defining diatom distribution. Most probably, the other types distributions are more strongly affected by other parameters applied in the model syntax. It is clear that the typical diatom parameterizations are not needed to a type to correlate with diatom OTUs, questioning the foundations of diatom classical parameterizations.

To investigate which values of the parameterized phytoplanktonic uptake and growth gives more probability to a type to correlate with diatom OTUs distribution, I looked into the distribution of types parametrizations weighted over the corresponding number of significant correlations (Fig. 6.8). Comparing to the parametrization of all the types (in red), diatoms simulating types (in blue) have higher maximum growth rates (PCMAX) and higher half saturation for growth over nitrate (KSATNO3). However, if we weight the parametrization of types over the number of times they significantly correlate with a diatom OTU, the distribution (in yellow) is more similar to the general parametrization of all the types.

This finding is supportive of the above observation that found significant correlations of diatom OTUs with all the types, independently by the

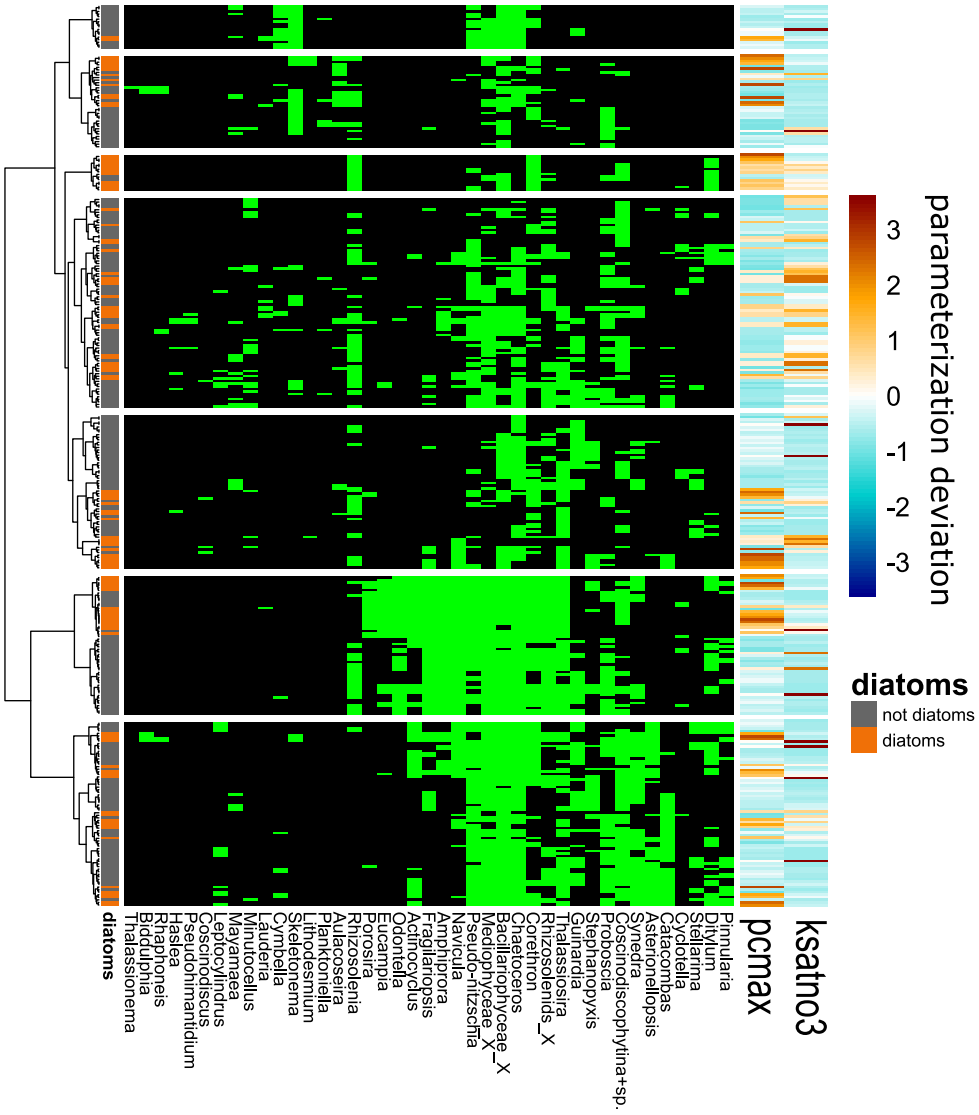

Fig. 6.7: On the left the same heatmap shown in Fig. 6.6 with a different color scale: where it is present at least a correlation, independently by the size fraction, between the OTUs of each diatom genus and a modeled phytoplankton type the cell is colored in green. If there is no correlation between any OTUs belonging to the diatom genus and the corresponding modeled type the cell is colored in black. On the left phytoplanktonic types have been annotated according to the fact that they simulate diatom organism or other phytoplanktonic taxa. On the right, a second heatmap shows the parametrization used to model the phytoplanktonic types which correspond to each row of the two heatmaps. The two parameters taken into account, PCMAX and KSATNO3, have been centered and scaled on every column.

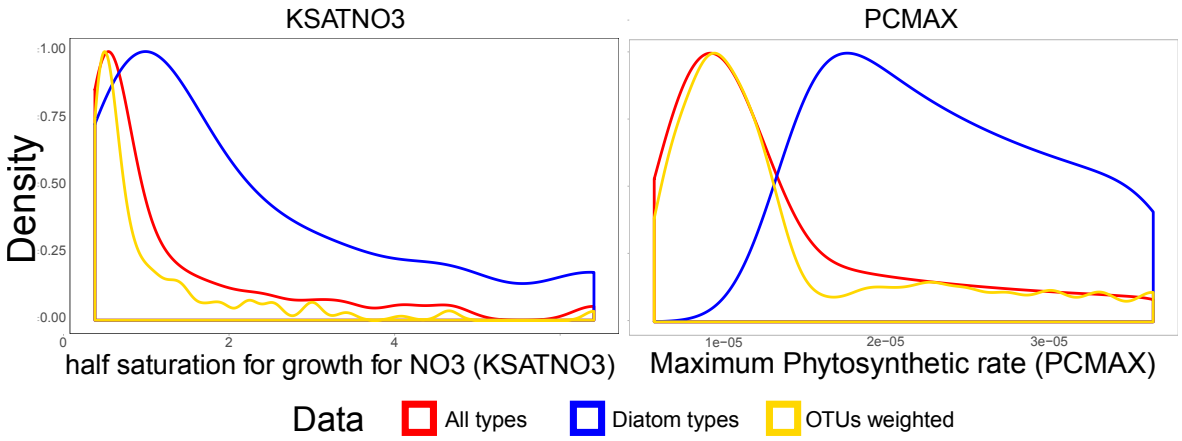

Fig. 6.8: Smoothed scaled density estimates for two parameters application across three size classes. Respectively in red and blue there is the density distribution of the applied parametrization of the parameter over the 350 phytoplanktonic type and the subset of 110 diatom simulating types included in the model. In yellow the parameters of each type are weighted over the number of significant correlation found for every type together with a diatom OTU. The plot describes only the size class 20-180 μm size class, but the other two size classes, not shown (5-20 μm; 180-2000 μm), exhibit strongly similar distributions.

fact that they are simulating diatoms or other phytoplankton like coccolithophores, dinoflagellates or other eukaryotes. Consequently, OTUs correlate to types independently by their parameterization, suggesting a wider range of parametrization needed for diatom simulating types.

Among the opportunities offered by the association of diatom OTUs to the model types there is the potential to characterize OTUs with the physiological characteristics used to define types. Virtually, this could be a very powerful means to characterize all of the unculturable taxa that until now were impossible to physiologically describe. Through this chapter I was not able to unequivocally associate OTUs to types and thus the assessment of OTUs PCMAX and KSATNO3 parameterizations is herein presented taking into account all the limits it has. Following I will present you few examples to test ecological and physiological question on diatoms through the potential of this association. The OTU parameterization was calculated as the median maximum growth rate (PCMAX) and the median of the half saturation for growth for nitrate (KSATNO3) of the phytoplankton types to which it correlates.

OTUs parameterization resulted in slightly different OTUs parameterization according to the size they belong to (Fig. 6.9). These results depict clearly higher maximum growth rates the higher the size class, in contrast to what we could expect. The relationship between diatom size and growth rate is supposedly influenced by many other elements such as for example temperature (Montagnes and Franklin, 2001), iron and light (Sunda et al., 1997). However, the interspecific variability of maximum growth rate has identified the cell size as one of the main controllers, with the smallest cells having the higher maximum growth rates (Sarthou et al., 2005; and the references therein) discordant with the results herein obtained. More recently, Marañon et al. (2013) proved experimentally that maximum growth rate peaked at intermediate cell sizes, however the only significant difference between size classes I observed was between big (180-2000 μm) and small (5-20 μm) diatoms, being both not significantly different from the intermediate sizes (20-180 μm).

Concerning the half saturation constant for growth over nitrate, smaller cells are expected to be more proficient in nutrient uptake, exhibiting smaller half-saturation constant than large cells (Eppley et al., 1969). The OTUs parameterization partially follows what have been previously described in the literature, having big and medium cell higher than the small ones. However big cells are not significantly different from the small ones.

Conclusively, OTUs parameterization depicts ambiguous results compared to what is known on diatom physiology across different size classes, but, again, this was somehow expected by the important limitation of this technical exercise.

Within the extensive diversity of diatoms we range from ubiquitous species to very rare species. My hypothesis is that while ubiquitous species are not specialized to any habitat but survive in all the communities without dominating, less ubiquitous species can be highly specialized to certain conditions,

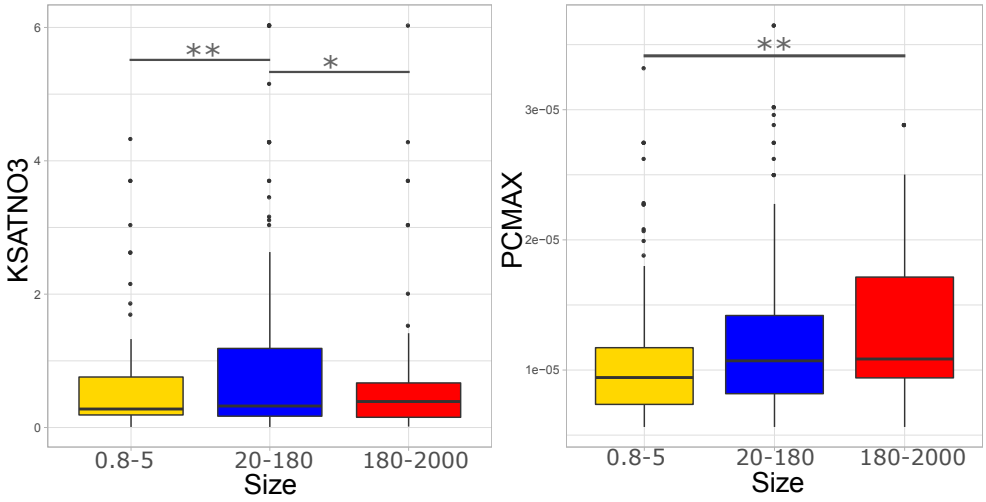

Fig. 6.9: Boxplot of OTUs associated KSATNO3 and PCMAX according to the size fraction. The mean of the distribution of both variables was compared among size fraction through pairwise t tests. Statistical significance difference between size classes is annotated as follow: * = p-value$\leq$0.05 and **=p-value$\leq$0.01.

and can develop periodic blooms. This basically reflects the tradeoff between capability to survive in a wide range of conditions and to be the fastest ones when nutrient pulses arrive. To investigate this relationship between ubiquity and the ability to form blooms, the ubiquity of each OTU was compared to the same OTU estimated PCMAX (Fig. 6.10). As previously noticed, the range of PCMAX of the OTUs is way wider than the one parameterized for the diatom-simulating phytoplanktonic types, in particular with lower values for the observed data compared to the modeled one. Nevertheless, what is noteworthy is PCMAX distribution across ubiquity: while most of the OTUs, independently by their ubiquity, have low maximum growth rate, the ones attaining higher PCMAX estimation, and therefore the ones prone to develop blooms, exhibit a general low ubiquity, suggesting a higher specialization to local environmental conditions.

A further information which can be provided by the model types' parameterization applied to OTU is a functional estimation of the different uses of nitrate by the OTUs (Fig. 6.11). The median KSATNO3 of OTUs, once classified in categories, can tell us the diversity of N utilization within the community if we refers to the presence absence of OTUs within the metabarcoding dataset.

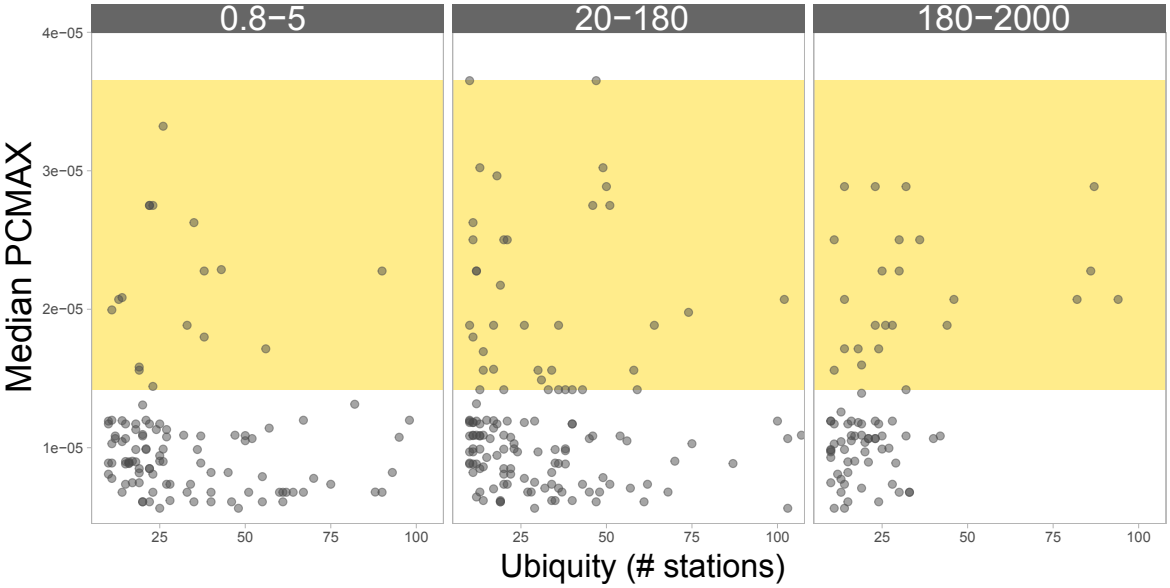

Fig. 6.10: Scatterplot of the relationship between the ubiquity of the OTUs expressed in number of *Tara* Oceans stations where it has been found present and the PCMAX derived for each OTUs. OTUs' PCMAX are computed as the median of PCMAX of the phytoplankton types correlating to the same OTU in the corresponding size. The yellow background corresponds to the range of PCMAX values applied in the model to parameterize diatom-simulating types.

This information can be though as a functional richness similar to the one we estimated upon N transporter gene families (chapter 4), as completely based on the use of N. There are strong limitations over this exercise, due to the not univocal correspondences between diatom and OTUs: indeed each OTUs is associated to types characterized by a wide range of KSATNO3 parameterizations (Fig. 6.12). Notwithstanding, similarly to what I found for the first approach (chapter 5) I observe a strong size effect on the richness distribution, discriminating peculiarly class 0.8-5 μm from the bigger classes. Overall, we observe higher diversity in the smaller classes and as we move to bigger classes the range of diversities observed goes down. There is no data of the Arctic stations for the small class, but for the bigger classes a latitudinal trends is observable with a peak, shared by all the classes, at the equator.

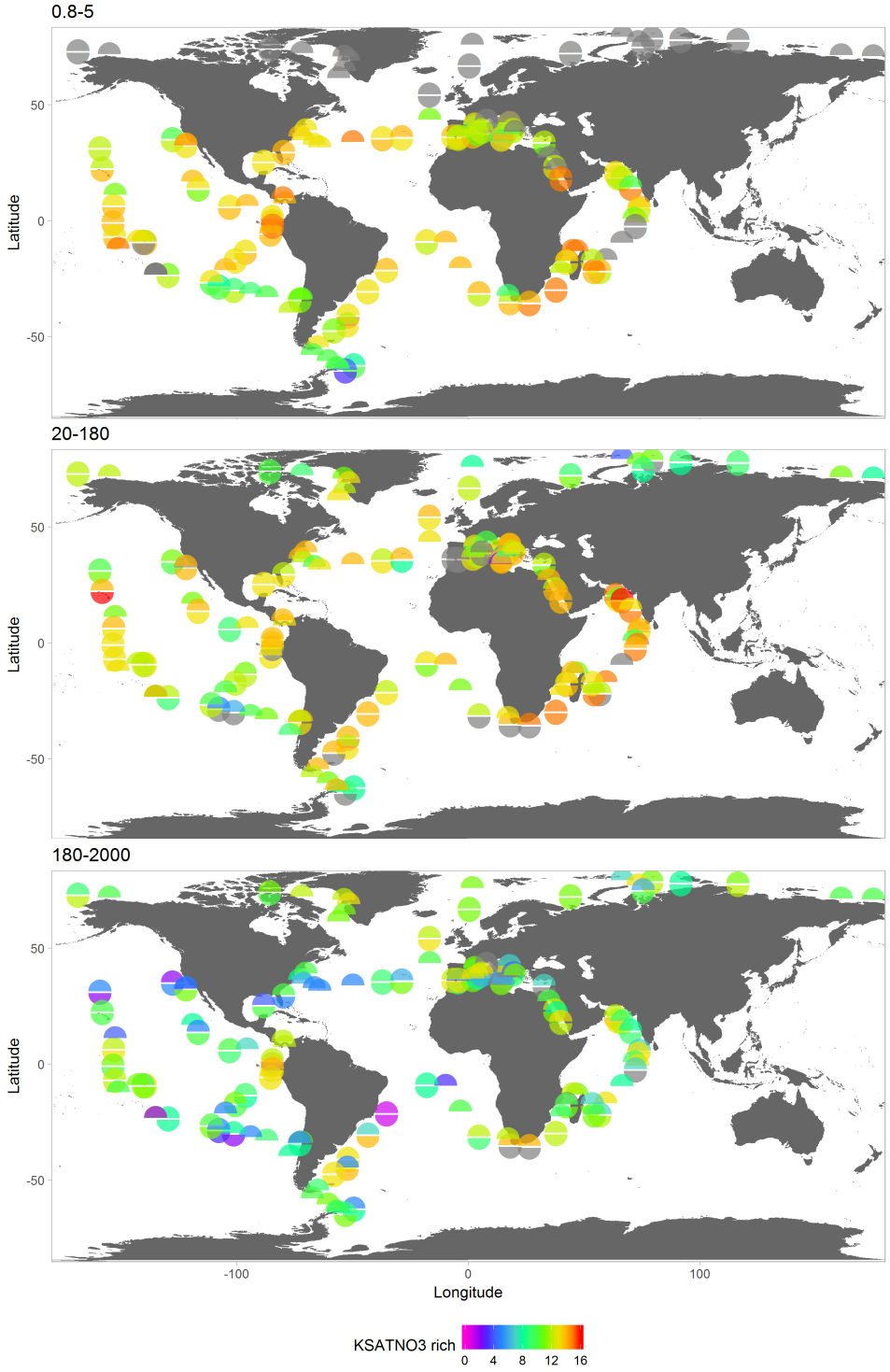

Fig. 6.11: Functional richness based on the different parametrization of KSATNO3 applied to OTUs. The median KSATNO3 was obtained per each OTU measuring the median of the parameterized KSATNO3 of the phytoplankton types correlating to the same OTU in the corresponding size. The same median KSATNO3 was then classified in 7 categories and the richness was estimated to be the number of different categories present in each sample. The top portion of each circle represents samples collected at the surface and the bottom portion represents the DCM (stations missing metatranscriptome data for one of the two depths are drawn as half circles).

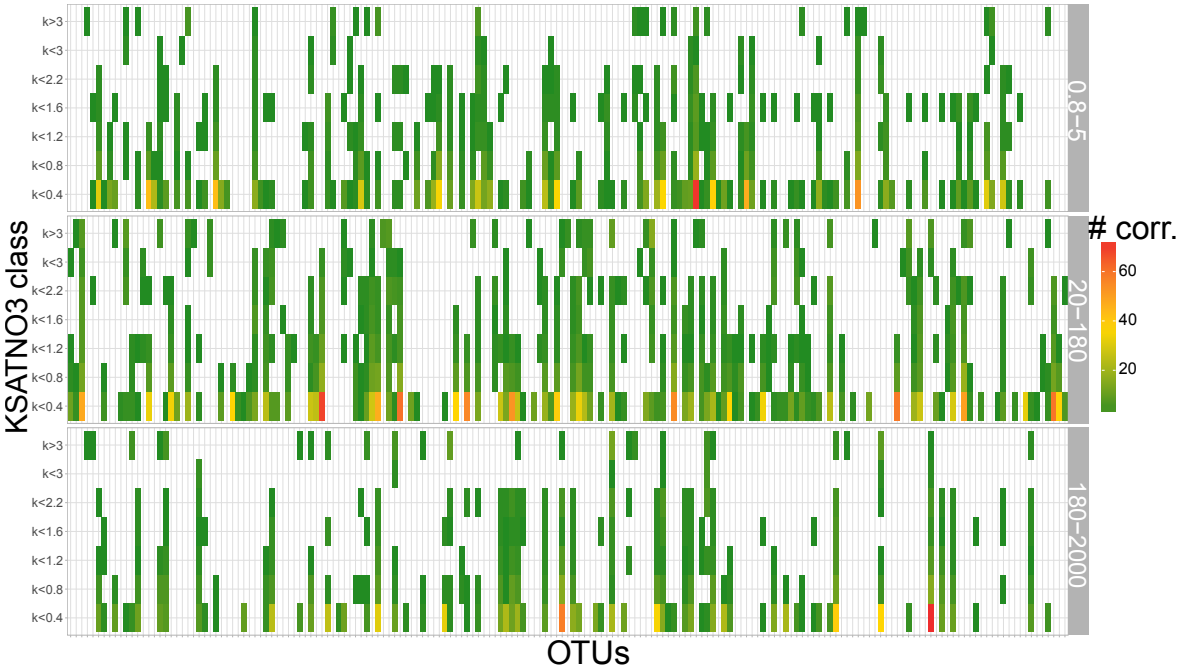

Fig. 6.12: Heatmap exhibiting the number of correlations of the diatom OTUs together with the phytoplanktonic types of the model, classed according to their KSATNO3 parameter.

6.5 Conclusions

Integrating sophisticated data such as metabarcoding and model outputs should not be given for granted. Within the present chapter I present a very preliminary exercise toward this direction.

Generally, PFT models include a classification of PFTs according to the scientific question addressed (Falkowski et al., 1998; Bouman et al., 2003, see chapter 1.1.3) but the resulting number of phytoplankton functional types included is usually very limited. As for example, the model PISCES (Aumont et al., 2015) includes two PFTs, silicifying diatoms and nanophytoplankton, NEMURO (Kishi et al., 2007) has two PFTs as well, diatoms and nanophytoplankton, PlankTOM5 (Buitenhuis et al., 2013) models three PFTs, diatoms, nanophytoplankton and calcifying coccolithophores, and CCSM-BEC (Moore et al., 2004) represents three PFTs as well (diatoms, nanophytoplankton and diazotrophs as nitrogen fixers). Six phytoplankton types were included in Le

Quéré et al. (2005) DGOM model: pico-autotrophs, phytoplankton N2-fixers, phytoplankton calcifiers, phytoplankton DMS-producers, phytoplankton silicifiers as diatoms and mixed phytoplankton as dinoflagellates. Of note, the number of PFTs is usually very restricted and diatoms are gathered within a single PFT, the only one requiring silicic acid. The same Le Quéré et al. (2005) addressed the need to further subdivide these PFTs, need limited only by the lack of sufficient information to further parameterize subclassifications.

The main conceptual problem in keeping large classifications (i.e., small number of classes) lies in the characterizing trait of each phytoplanktonic type. Indeed, each PFT is characterized by a specific set of physiological traits. However, if these traits values are obtained through lab experiments, they have been deduced by single culturable species, which are not necessarily representative of the whole PFT. Nevertheless, even if these traits parameterizations are based on empirical observation, the applied values may not represent correctly the responses range (Le Quere et al., 2005). A step further was done by Follows et al. (2007) with their emergent adaptive model within which the authors modeled four functional classes of phytoplankton (i.e., *Prochlorococcus* analogs, small photo-autotrophs, diatoms and large phytoplankton), subdivided in many types (in the publication 78) whose physiological characteristics were stochastically determined. The model is adaptive as it starts its simulation with all the types being everywhere with the same biomass, but, as the model is integrated over 10 years, specific communities and distributions arise. Further work on this model has been done by several authors, among those there is the model I worked with within this chapter: based on Dutkiewicz et al. (2015). Herein they resolved nine phytoplankton "functional" types: diatoms analogues, other large eukaryotes, coccolithophores, picoeukaryotes, *Synechococcus*, high- and lowlight *Prochlorococcus*, nitrogen-fixing *Trichodesmium*, and unicellular diazotrophs. Every functional class is further subdivided in physiologically differently regulated units, with stochastics regulations of traits counting a final number of 350 phytoplanktonic types.

The modeling field is going toward complexity, increasing more and more the number of types, whose only limitation is the difficulty in evaluating the parameters controlling them, because of the extremely limited knowledge from laboratory cultures (Le Quere et al., 2005; Follows et al., 2007). The aim of this chapter, to compare phytoplanktonic types from a mathematical model to the metabarcoding OTUs of *Tara* Oceans, is only possible thanks to this newly reached complexity and finer scale modeling of phytoplanktonic types. It is the first time a model includes 350 different phytoplankton functional types, and with these numbers it should be now possible to compare the units to lower taxonomic ranks units.

The exercise is undoubtedly preliminary and extremely simplistic being based only on correlations, but even with such an attempt of comparison between the two datasets I could notice different features. First, I found no preferential correlation of diatom OTUs together with diatom simulating types. Diatom OTUs correlated indeed with all the types indiscriminately. Diatom simulating PFTs are the only types which need silica to growth. Compared to the other types they are usually modeled with the highest growth rates (Irwin et al., 2006), they have the higher Chl α : C (MacIntyre et al., 2002) and they have the highest half saturation constant for growth over all the micronutrients (Dutkiewicz et al., 2015). Therefore, diatoms analog types are somewhat an extreme in terms of physiological characterizations. The fact that diatom OTUs correlated with all the types is suggestive of a likely broader range of diatom physiological descriptions to include the incredible wide diversity of diatoms. In particular, while diatoms are characterized by high growth rate and high KSATNO3, this study highlighted that different species of diatoms may have lower growth rate or KSATNO3. One possibility for completing the exercise would be to develop finer ways to associate OTUs and types, for example using machine learning approaches and other tools from Artificial Intelligence.

A second observation is that the significant correspondences between OTUs and types were not univocal. To every type, multiple OTUs were found significantly correlated, which is expected because types are just aggregation of species based on their functional role stipulated by similar physiologies. More peculiar is the correspondence of more diatom types to the same diatom OTU, indicating similar patterns, even if only locally, between different types. This is both a limit of the methodological approach herein undertaken and a limit of the model. The latter indeed may ends up building redundancy between types distribution, modeling such a high number of types. Is this an error? Not forcely. One explanation may be functional equivalence of different types. Indeed, looking at the heatmap of distances between types' distributions (Fig 6.2) it is clear that groups of types are characterized by similar patterns, but diatoms are widely distributed in the ordination dendrogram, indicating that packages of types do not correspond to similar physiology behaving in the same matter but rather co-occurrence of differently parameterized types.

It will be fundamental in future studies to better integrate omic- and modeling- derived findings. This integration has the potentiality to provide a clear picture of the taxonomic tag of the main players of planktonic communities as well as of the processes ruling them. Obviously, a model complex enough to simulate all the plankton system is still unimaginable considering our actual knowledge of the system but one step further in complexity will be now accessible thanks to omic data.

7 Thesis summary and outlook

7.1 Thesis scope and main results

The general aim of my thesis was to investigate the diatom diversity information provided by omic data, exploiting the availability of the unprecedented amount of data sampled by the *Tara* Oceans expedition. So what are the achievements of this thesis? These latter can be divided in the methodologic and scientific challenges addressed by this work.

Concerning the methodological aspect of my thesis, I found very challenging to exploit and make the most of omic data to investigate diatom diversity. Technologies to retrieve omic data are around since the 90s but only in the last decade large datasets have become accessible only thanks to their cost drop. Given the youth of this field, the scientific community has still not assessed standardized procedures to deal with this kind of data. As I investigated both taxonomic and functional diversity of diatom several questions arose on the better way to retrieve these information. Taxonomic diversity of diatoms was estimated using a 18S metabarcoding improved by an in-house developed filtering approach based on the integration of morphology-based identifications (chapter 2). I proved that, excluding only the 0.35% of the cumulative abundance of rare OTUs, the information retrieved by the two measures reconciliated. Rather than applying marked filtering pipelines such as the common 1% threshold on the taxa relative abundance (e.g., Vallina et al., 2014a) and focus only on the more abundant species, the approach I developed allows to maintain the rare component of the system, supposedly

excluding only artefacts or evolutionary close subpopulations or strains of more abundant OTUs.

However, the most important methodologic advancement proposed by this thesis is the pipeline of analysis developed to assess diatom functional diversity over metatranscriptomic data (chapters 3, 4 and 5). There are two different conceptual points of view on how to measure functional diversity: from the system point of view, taking into account all the possible functions within it, where one usually just classifies taxa belonging to each function. On the other hand, as taking all the functions into account is hardly feasible, ecologists often focus on one (or more) specific function and identify and characterize the different ways displayed by the organisms to complete it. Following this latter approach, I developed a pipeline to characterize one aspect of the functional diversity: the N uptake. The idea behind all this is that diatoms adopted specific evolutionary solutions to respond to different environmental systems and to the corresponding biotic interactions context. That is that the genetic diversity of these families could correspond to the functional diversity of the diatoms owning them. Consequently the herein designed pipeline is based on the definition of functional units over phylogenetic clusters of key gene families (chapter 3). The innovation of this choice is powerful as functional units are built over strong evolutionary hypothesis beneath, unlike a large part of functional studies, which define functionality over the resulting distribution of species. Rather than starting by the consequences of functionality (i.e., the distribution) my approach finds its strength and conceptual rigors in starting by the evolutionary point of view (i.e., phylogenetic closeness). Following the definition of functional units the pipeline contemplates multivariate and machine learning approaches to investigate both the putative functional diversity of diatoms across the oceans but also to identify the functional role of the units themselves. This downstream set of analysis is supportive of the differential functionalization of phylogenetic clades, as I found a differential use of the latter according to different environmental conditions. Nevertheless, it is not possible to define the actual functional role of different clades as biological

information on the genes constituting clades are missing and only laboratory experiments could prove it.

From a scientific point of view several aspects of diatom diversity were carefully investigated. Following the methodological challenge to obtain taxonomic and functional diversity of diatoms from omic data I discerned their global distributions looking for specific patterns and for the responsible environmental drivers beneath. The obtained taxonomic and functional diversity resulted in depicting patterns of global diversity that added upon previous modeling simulations. Interestingly, the functional diversity derived over di-*AMT1* phylogeny, was the one better validating numerical modeling simulations such as Vallina et al. (Vallina et al., 2014a) one, able to detect the major predicted hotspots in the equatorial Pacific regions as well as in the mid-temperate areas (chapter 4).

In the presented thesis I have provided new information on the environmental cues triggering diatom taxonomic hotspots formation as well. According to my analysis (chapter 2) the latter are explainable as consequences of one of two processes: micronutrient availability and hydrodynamic processes, as well as by the combination of the two. I also showed how, according to the local environmental context, diatom taxonomic richness can be differentially controlled by several variables. One variable rather than the other may be determinant in shaping diversity according to the region, unveiling the very complex position of diatom within the planktonic community.

Regarding diatom functional diversity I proved N transporter gene families to be excellent markers of diatom functional diversity. I provided sufficient evidence of the power of the phylogenies of these genes in characterizing diatoms functionality not only according to their answer to different N sources and different N sources availabilities, but also to iron availability and in relation to the diatom cell size, a classical functional trait (chapter 5). Functional

units built over these families allowed me to compute putative diatom functional richness (chapter 4). A biogeography can be discerned by both the distribution of the units (chapter 4) as well as by their mRNA abundance modulation (chapter 5). A specific shift in functionality terms has been observed with depth, comparing surface samples to the corresponding deep chlorophyll maximum depth ones. This shift differs between the two gene families analyzed, highlighting the precise reaction of diatoms to different N sources availabilities. I have illustrated as well a comparison of the use of putative functional units to N metabolism prokaryote modules at the two sampling depths, demonstrating strong correlations between the two, suggestive of processes of cooperation and/or competition between prokaryote and diatoms. Ammonium transporters proved to be a better marker to include the information of the relationship of diatoms together with N fixing bacteria.

Furthermore, in this thesis I have illustrated the potentiality and interest of integrating mathematical modeling and omics approaches for the understanding of phytoplankton diversity. From this preliminary analysis I demonstrated how the actual parametrization of phytoplankton functional types simulating diatoms may be too narrow to cover all the diversity of diatoms. While in the models diatoms are usually described as strongly opportunistic species, able to rapidly take advantage of nutrients when available, they may not all be portrayed by these characteristics. They are not all so efficient in nutrient uptake and assimilation, and neither they all have this high growth rate that usually distinguish them in the model syntax. Even if the correspondence between diatom OTUs and the modeled phytoplankton types was not univocal, and thus the inference of physiological characteristics of OTUs from the model parameterization is still not strongly significant, I wanted to develop two more exercises to show the potentiality of this association. I was able to show that the less ubiquitous OTUs, and hence, conceptually, the more specialized ones, are the ones able to form blooms thanks to their high maximum growth rate. Furthermore, looking at the different values of

K for nitrate utilization I exhibited patterns of functional diversity of diatoms based on their N utilization, a completely different approach to measure a conceptually equivalent diversity measure as the one derived in this thesis by meta-omic data. Indeed, the potential of omic and modeling is great, and further efforts are needed to strengthen this association as it could virtually give access to the physiological characterization of all the unculturable taxa within the plankton.

7.2 Thesis summary

The present thesis is structured to cover three main subjects: i) the study of diatom taxonomic diversity; ii) the study of diatom functional diversity and iii) the integration and omic and numerical modeling to improve the understanding of the planktonic system.

In the first chapter all the background information on the main topics addressed by this thesis are provided to better understand the present work and locate it in the right scientific context. It has been presented what is known in relation to phytoplankton and specifically to diatom diversity, what are the processes behind it, analyzing in particular the possible drivers of phytoplankton species coexistence. I introduced an overview of how phytoplankton is commonly modeled, what is the expected and observed biogeography of diatoms and how these latter in particular take part to the main ocean biogeochemical cycles. Furthermore, I resumed what is known on diatom nitrogen metabolism, focusing on the regulation of specific marker genes within this same metabolism. To conclude, I presented the recent omic revolution, illustrating all the potential of these methods to improve our understanding of phytoplankton dynamics.

In chapter 2 I investigated taxonomic diatom diversity integrating metabarcoding and morphology-based counting. I filtered the metabarcoding of the rarest OTUs to reconcile at best the richness information retrieved from the genomic material and from the microscopy counting exercise. As the filtering procedure is the result of a herein developed pipeline, I investigated the effect of this filtering. The rarest OTUs excluded by the process may be strains or cryptic species evolutionary close to a more abundant OTU not excluded by the filtering process. A second explanation is that filtered OTUs may be spurious caused by artifact. The percentage of filtered OTUs is highly variable, strongly dependent on geography, specifically increasing with latitude. Consequently, at the poles, where diatoms are relatively more abundant, samples have been affected the most by the filtering procedure. The relation with the abundance of diatoms could be suggestive of artifacts as the cause of filtered rare OTUs. However, phylogenetic diversity analysis revealed that filtered OTUs may be originated by rare strains, as there is a linear relationship between the loss of phylogenetic diversity produced by the filtering procedure and the percentage of OTUs excluded by the same. Finally, I depicted the patterns of diatom taxonomic diversity and focused on the environmental driver of this diversity. Different areas are more controlled by one environmental variable over the others, but, generally, all the variables taken into account (i.e., micronutrients availability, water dynamics and temperature or chlorophyll α concentration) had a fundamental role shaping diatom taxonomic diversity. The formation of hotspots was demonstrated to be linked to lateral transfer, which can lead to the overlap of neighbour communities, and/or to the presence of micronutrient availability. This two processes can contribute together or one of the two may be the main hotspot cause, according to the specific hotspot.

The following three chapters examined diatom functional diversity across the *Tara* Oceans metagenomic and metatrascriptomic datasets. In chapter 3 I built putative functional units of diatoms based on phylogenetic clustering. The marker gene families selected to run this exercise are two N transporter

gene families. The sequences corresponding to these families found on the metatranscriptome of *Tara* Oceans have generally low ubiquity but they well represent diatom taxonomic diversity, being spread across the genera in expected relative abundances. This study demonstrated the limits and weakness of the metagenomic dataset when looking at genes within specific gene families. Consequently the whole study was based on the metatranscriptome, in particular on the presence-absence and distribution of the established functional units across all the available *Tara* Oceans samples along three size fractions (i.e., 0.8-5 μm, 20-180 μm, 180-2,000 μm micrometers).

The following chapter (chapter 4) focused on the distribution of the designed putative functional units based over the two gene families. The functional classification applied was assessed as successful because of the pattern of functional diversity observed, which mirrored and validated what expected by modeling approaches. The two gene families resulted in similar patterns even if studies on the expression modulation of the two often resulted in different regulations. This is suggestive of the quality of the outcome of the methodological pipeline to design functional units. Comparing the taxonomic richness observed in chapter 2 to the putative functional richness based on the two gene families gave interesting insights on the two different diversity information. In particular high taxonomic richness corresponds to medium functional richness, highlighting the presence of ecologically redundant species within the same environment in case of taxonomic hotspots.

Finally, in chapter 5 I investigated the different functional roles played by the different clades. To do so the environmental drivers behind the single functional units were thoroughly investigated. I work with a double information: the presence/ absence of a functional unit and its mRNA level in the samples. This means that there are two levels of modulation: a first one is the distribution of clades and a second one is the modulation in terms of mRNA abundances where they are present. Exploiting these information I

found a strong regionalization of the presence and, following, also of the abundance of my units. Units are not only preferentially used in specific geographical regions but also in specific fraction sizes, delineating deeply different functional roles of the same. A further differential use of the units was observed on the comparison between surface and DCM samples, highlighting metabolism shifts with depth, to respond to very different external conditions. Parallelisms were found with the prokaryotic functional modules linked to nitrogen metabolisms. This is suggestive of a possible interaction between the two planktonic compartments, which could be of cross-feeding (i.e., bacteria would help diatoms by remineralizing N and/or fixing N, transforming it in an accessible form to diatoms) or competition (i.e., bacteria and diatom would compete for the same sources of N). Finally, machine learning approaches allowed insights over the specific functional role of each unit looking at the optimal environmental condition for their presence and for their higher abundance. Specifically, functional units were differentiated by nitrogen sources, iron availability and temperature. A simple prediction analysis allowed me to individuate the weaker functional units in case of a temperature increase, delineating how few functional units could be dramatically affected by such a change.

To conclude the thesis, in chapter 6 I presented a simple exercise of integration of a mathematical model including 350 phytoplankton functional types with the metabarcoding from *Tara* Oceans, already introduced in chapter 2. Although the analyses are still very preliminary it is clear that it is possible to find connections between diatom OTUs and the modeled phytoplankton types. The advantages are manifold as through this integration we could give a specific taxonomic name to the PFT of the models but in the same time we could also obtain physiological information on uncultured species. What emerged from this analysis is the comparability of diatom OTUs to all the phytoplankton functional types and not only to the ones simulating diatoms, showing how diatoms could be parameterized with wider physiological ranges

of abilities, in line with very recent suggestions (Kemp and Villareal, 2018). I parameterized diatom OTUs physiology using as a proxy the parameterization of the model type they correlate too. Even if this OTUs characterization is still very weak, because of the several significant correlations found for each OTUs, this step allowed me to proceed with few conceptual ecological exercises, in order to highlight the potential of a correct combination between the two datasets.

7.3 General considerations and future perspectives

Dealing with metaomic dataset is not easy. This kind of information is incredibly promising (Buttigieg et al., 2018) but its analysis is still in complete evolution. Recently, the international group on Earth Observations Biodiversity Observation Network (GEO BON) highlighted among the challenges and opportunity of marine biology the need for the oceanographic community to integrate and develop technologies in particular to deal with the omics data (Muller-Karger et al., 2018). There are still no standard approaches to have insights over this data and as this novelty opens the doors to extensive possibilities, the scientific communities is still struggling looking for the best approaches.

While I believe the methodological frameworks I propose here are useful, they are subject to several limitations. First of all the entire study is to be considered reference based. The identification of diatom OTUs and the identification of diatom N transporter genes is completely based on the direct comparison to diatom references. We have to consider that the number of existing diatom species is estimated to tens to hundreds of thousands (Guiry, 2012; Mann and Vanormelingen, 2013), while the number of assembled genomes up to now is of only 7 species (*Thalassiosira pseudonana*, *Thalassiosira oceanica*,

Phaeodactylum tricornutum, *Fragilariopsis cylindrus*, *Pseudo-nitzschia multiseries*, *Fistulifera solaris*, and *Synedra acus*). Higher number of species are covered by the transcriptomes, which are around 100 thanks to the Marine Microbial Eukaryote Transcriptome Sequencing Project (Keeling et al., 2014), but still the reference catalogue is very far to sufficiently represent diatom diversity. The building of a more complete reference catalogue for diatoms will have a big impact on functional and comparative genomic studies, giving access to finer quality searches within the omic samples. Only a very restricted portion of diatom diversity has been subject of genomic studies. *Tara* Oceans metabarcoding approach counted 63,371 unique diatom-derived barcodes, annotated to 4,748 taxa (Malviya et al., 2016), including a very large portion of yet unidentifiable species. It is therefore fundamental to keep on improving and growing the reference dataset, including those still uncultured taxa (Tirichine et al., 2017). Within this contest, single cell sequencing is emerging as a fundamental step to gain knowledge on the genome structure and function of a species (Bhattacharya et al., 2012; Roy et al., 2014). An important effort should be taken in this direction to fill the information gap on unculturable species, which are the majority of diatoms (Tirichine et al., 2017). This technology will allow not only to minimize the reference bias produced by catalogue reference representing only culturable species, but also to analyze diatoms with their interacting organisms from symbionts to parasites (Riemann et al., 2000; Lima-Mendez et al., 2015). Gathering this information would help better locate diatoms within the community and the food webs.

A second important limitation of this study was the lack of saturation within the metagenomic dataset. The functional part of the thesis was indeed solely based on the metatranscriptomic data only because of this limit. Ideally, I would have took advantage of metagenomic dataset to assess the distribution of genes and to normalize their observed expression in the metatranscriptome. Nevertheless, I get around the problem proposing alternative

methodological choices. Saturation in this kind of data is very hard to obtain (chapter 1.4), and this becomes rapidly evident when you focus on specific genes. Further studies should improve our understanding of the real coverage of the metagenomic dataset, expressed as the fraction of the metagenome represented on the dataset (Rodriguez-R and Konstantinidis, 2014). But we also need better control on the sequencing depth in order to properly sequence the samples to reach saturation. As these new sequencing techniques are becoming more and more affordable, the moment has arrived to require more advanced protocols including biological and technical replicates. To accurately explore omic dataset at the gene level, physiological and expression regulation studies are now indispensable. This experimental work is indeed fundamental to develop functional genomic, but in the diatoms case it could be an even more discriminating information considering the recent assessment of diatoms as natural hybrids (Rynearson and Armbrust, 2005; Casteleyn et al., 2009). Having a clear idea of the theoretical framework behind physiological answers to external cues could allow us to properly understand the functioning of basic processes at the wider scale provided by recent global genomic surveys as *Tara* Oceans. There is now the need to identify specific biologically-based life traits to correctly build phytoplankton functional units in the models. These traits should focus on underlying mechanisms of cell activity and physiology, characterizing phytoplankton dynamics at a population level (Allen and Polimene, 2011). In my thesis I designed functional traits over N transporter genes (chapter 3), one of the encountered limitation was the scarce previous knowledge on the functional role of single genes: transcriptomic studies were done only on few species and on a very limited set of conditions (chapter 6). A deeper knowledge of the regulatory system of these genes could have eased the functionality assessment of the units. New bioinformatic tools could improve these kind of studies, allowing for example to detect cases of mono-allelic gene expression or allele exclusion, to fully understand their role in transcriptional regulation (Tirichine et al., 2017). An even finer analysis would include the use of Single Nucleotide Polymorphisms (SNP), or other genomic markers, for

the discrimination of different populations within meta-omic data like the *Tara* Oceans one (Arif et al., Submitted). This approach could be virtually informative of the evolution and functionalization of genes within the same gene family. Following this idea, a current PhD project is now born as a follow up of the present work with the aim of further investigate the evolution, distribution and functionalization of genes within the same datasets here utilized, thanks to SNPs detection.

One strong assumption of my thesis was the definition of functional units over only one gene family. This is however to be considered as a first step in defining a functionality strictly linked to N metabolism, and as this metabolism is strictly linked to several other activities of the cell, the functional diversity I obtain could be a concrete representative of diatom functional diversity. A further step should be made by the inclusion of other key gene families in the functional characterization, in order to include more than one biological trait in the definition. One strong point of the *Tara* Oceans sampling framework is its coverage of all the Oceans, representing in varying degrees most of the different hydrological structures present in the oceans. However, one strong limitation of *Tara* Oceans is the punctual sampling: we are just observing a one-time situation of the communities and of the hydrodynamics while the majority of the stations taken into account (excluded the equatorial samples) are characterized by strong seasonal changes, in both biotic and abiotic terms. The next step of the meta-omic sampling effort should be now focused on time series. Sampling the same location during the year, such as well characterized ecosystem (e.g., LTER), or sampling the dynamics through specific processes like bloom wax and waning, should be the aim of the new-born genomic-enabled observatories.

Moreover, through my thesis (chapter 6) I illustrated the potentiality of integrating omic and modeling results for a better understanding of phytoplankton communities. Of course this process is permitted by a growing

refining of the phytoplankton functional units included in models. Generally, modeling approaches resolve phytoplankton diversity through few types characterized by highly simplified physiologies (e.g., Le Quere et al., 2005; Litchman et al., 2007) and within each type several units are distinguished by cell size or temperature preferences (Follows et al., 2007). This thesis promotes the improvement of phytoplankton physiology within the models, in order to improve their effectiveness in describing planktonic dynamics. The main limit encountered by the approach herein proposed in associating diatom OTUs together with model phytoplankton types is the not-univocal combination of the two classes. Finer approaches, based not on mere correlations but on the definition of the environmental niches of the types and of the OTUs followed by the comparison of the multivariate space occupied by OTUs and types niches could provide more robust associations between the two.

We are now in the era of continuous production of omic datasets, counting several global surveys of the most diverse microbial communities. Nevertheless, still a great effort from the scientific community is needed to fully exploit and understand this data. This thesis takes part to this omic revolution improving the use of omic data specifically for diatoms, one of the most important player of phytoplankton communities. The methodological pipelines here proposed are surely improvable, overcoming the limits listed in this chapter, and they can be applied to all the elements of the planktonic community. Moving toward a fine scale understanding of the planktonic dynamics will allow to better model them and finally to better predict their impact on biogeochemical cycles or the entire food webs, facing environmental changes such as the ones predicted for climate change. Omic data has unimaginable potential information and it's become relatively accessible in economic terms: there's nothing left to do for the scientific community to learn how to exploit it to the fullest.

Bibliography

Abad, D., A. Albaina, M. Aguirre, et al. (2016). „Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy“. In: *Marine Biology* 163.7, pp. 1–13 (cit. on p. 82).

Abrams, P. (1983). „The theory of limiting similarity“. In: *Annual Review of Ecology and Systematics* 14.1, pp. 359–376 (cit. on p. 13).

Adler, P. B., J. HilleRislambers, and J. M. Levine (2007). „A niche for neutrality“. In: *Ecology Letters* 10.2, pp. 95–104 (cit. on p. 13).

Aksnes, D. L. and J. K. Egge (1991). „A theoretical model for nutrient uptake in phytoplankton“. In: *Marine Ecology Progress Series* (cit. on p. 201).

Aksnes, D. L. and F. J. Cao (2011). „Inherent and apparent traits in microbial nutrient uptake“. In: *Marine Ecology Progress Series* (cit. on pp. 168, 201).

Alberti, A., J. Poulain, S. Engelen, et al. (2017). „Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition“. In: *Scientific Data* 4. arXiv: `arXiv:1208.5721` (cit. on pp. 36, 94).

Alexander, H., B. D. Jenkins, T. A. Rynearson, and S. T. Dyhrman (2015). „Metatranscriptome analyses indicate resource partitioning between diatoms in the field“. In: *Proceedings of the National Academy of Sciences* 112.17, E2182–E2190 (cit. on pp. 11, 12, 91, 92, 150, 152, 154).

Alipanah, L., J. Rohloff, P. Winge, A. M. Bones, and T. Brembu (2015). „Whole-cell response to nitrogen deprivation in the diatom Phaeodactylum tricornutum“. In: *Journal of Experimental Botany* 66.20, pp. 6281–6296 (cit. on pp. 92, 149, 151, 153, 163).

Alipanah, L., P. Winge, J. Rohloff, et al. (2018). „Molecular adaptations to phosphorus deprivation and comparison with nitrogen deprivation responses in the diatom Phaeodactylum tricornutum“. In: *PLoS ONE* 13.2, pp. 1–24 (cit. on pp. 150, 151, 153).

Allen, A. E., C. L. Dupont, M. Oborník, et al. (2011). „Evolution and metabolic significance of the urea cycle in photosynthetic diatoms“. In: *Nature* 473.7346, pp. 203–207 (cit. on pp. 26, 27, 150, 153).

Allen, A. E., A. Vardi, and C. Bowler (2006). „An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms“. In: *Current Opinion in Plant Biology* 9.3, pp. 264–273 (cit. on p. 27).

Allen, A. E., B. B. Ward, and B. Song (2005). „Characterization of diatom (Bacillariophyceae) nitrate reductase genes and their detection in marine phytoplankton communities“. In: *Journal of Phycology* (cit. on pp. 27, 91).

Allen, E. E. and J. F. Banfield (2005). „Community genomics in microbial ecology and evolution“. In: *Nature Reviews Microbiology* 3.6, pp. 489–498. arXiv: 2031 (cit. on p. 28).

Allen, J. I. and L. Polimene (2011). „Linking physiology to ecology: Towards a new generation of plankton models“. In: *Journal of Plankton Research* (cit. on p. 233).

Amato, A., G. Dell'Aquila, F. Musacchia, et al. (2017). „Marine diatoms change their gene expression profile when exposed to microscale turbulence under nutrient replete conditions“. In: *Scientific Reports* 7.1, pp. 1–11 (cit. on pp. 5, 150, 152, 154).

Amato, A., V. Sabatino, G. M. Nylund, et al. (2018). „Grazer-induced transcriptomic and metabolomic response of the chain-forming diatom Skeletonema marinoi“. In: *ISME Journal* 12.6, pp. 1594–1604 (cit. on pp. 5, 152, 154).

Amin, S. A., L. R. Hmelo, H. M. Van Tol, et al. (2015). „Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria“. In: *Nature* (cit. on p. 79).

Anderson, M. J. (2001). „A new method for non parametric multivariate analysis of variance“. In: *Austral ecology* 26.2001, pp. 32–46 (cit. on p. 135).

Anderson, M. J. (2006). „Distance-based tests for homogeneity of multivariate dispersions“. In: *Biometrics* 62.1, pp. 245–253 (cit. on pp. 129, 135).

Anderson, T. R. (2005). „Plankton functional type modelling: Running before we can walk?“ In: *Journal of Plankton Research* 27.11, pp. 1073–1081 (cit. on pp. 14, 15).

Andrade, S. L. and O. Einsle (2007). *The Amt/Mep/Rh family of ammonium transport proteins (Review)* (cit. on p. 104).

Anisimova, M. and O. Gascuel (2006). „Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative“. In: *Systematic Biology* 55.4, pp. 539–552 (cit. on p. 96).

Antia, N. J., C. D. McAllister, T. R. Parsons, K. Stephens, and J. D. Strickland (1963). „Further measurements of primary production using a large-volume plastic sphere“. In: *Limnology and Oceanography* (cit. on p. 170).

Arif, M., K. Sugier, D. Iudicone, et al. (Submitted). „Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and the Mediterranean Sea“. In: (cit. on p. 234).

Armbrust, E. V. (2009). „The life of diatoms in the world's oceans“. In: *Nature* 459.7244, pp. 185–192 (cit. on pp. 4, 21, 22).

Armbrust, E. V., J. A. Berges, C. Bowler, et al. (2004). „The genome of the diatom Thalassiosira Pseudonana: Ecology, evolution, and metabolism“. In: *Science* 306.5693, pp. 79–86. arXiv: 9809069v1 [arXiv:gr-qc] (cit. on pp. 25–27).

Ashworth, J., S. Coesel, A. Lee, et al. (2013). „Genome-wide diel growth state transitions in the diatom Thalassiosira pseudonana“. In: *Proceedings of the National Academy of Sciences* 110.18, pp. 7518–7523 (cit. on pp. 149, 150, 152).

Assmy, P., V. Smetacek, M. Montresor, et al. (2013). „Thick-shelled, grazer-protected diatoms decouple ocean carbon and silicon cycles in the iron-limited Antarctic Circumpolar Current“. In: *Proceedings of the National Academy of Sciences* (cit. on p. 22).

Aumont, O., C. Ethé, A. Tagliabue, L. Bopp, and M. Gehlen (2015). „PISCES-v2: An ocean biogeochemical model for carbon and ecosystem studies“. In: *Geoscientific Model Development* 8.8, pp. 2465–2513 (cit. on pp. 130, 219).

Baas-Becking, L. G. M. (1934). *Geobiologie of inleiding tot de milieukunde* (cit. on pp. 17, 18).

Baines, S. B., B. S. Twining, M. A. Brzezinski, et al. (2012). „Significant silicon accumulation by marine picocyanobacteria“. In: *Nature Geoscience*. arXiv: 9605103 [cs] (cit. on p. 22).

Barton, A. D., S. Dutkiewicz, G. Flierl, J. Bragg, and M. J. Follows (2010). „Patterns of Diversity in Marine Phytoplankton“. In: *Science* 327.5972, pp. 1509–1511 (cit. on pp. 12, 19, 47, 48, 79, 80, 83, 86, 125, 131, 140, 141, 144, 184).

Barton, A. D., A. J. Irwin, Z. V. Finkel, and C. A. Stock (2016). „Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities“. In: *Proceedings of the National Academy of Sciences* 113.11, pp. 2964–2969 (cit. on p. 16).

B-Béres, V., Á. Lukács, P. Török, et al. (2016). „Combined eco-morphological functional groups are reliable indicators of colonisation processes of benthic diatom assemblages in a lowland stream“. In: *Ecological Indicators* (cit. on pp. 88, 90).

B-Béres, V., P. Török, Z. Kókai, et al. (2017). „Ecological background of diatom functional groups: Comparability of classification systems“. In: *Ecological Indicators* 82.July, pp. 183–188 (cit. on p. 90).

Behl, S., A. Donval, and H. Stibor (2011). „The relative importance of species diversity and functional group diversity on carbon uptake in phytoplankton communities“. In: *Limnology and Oceanography* (cit. on p. 8).

Beltrán-Heredia, E., D. L. Aksnes, and F. J. Cao (2017). „Phytoplankton size scaling with nutrient concentration“. In: *Marine Ecology Progress Series* (cit. on p. 201).

Beman, J. M., J. A. Steele, and J. A. Fuhrman (2011). „Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal California“. In: *ISME Journal* (cit. on p. 33).

Bender, S. J., C. A. Durkin, C. T. Berthiaume, R. L. Morales, and E. V. Armbrust (2014). „Transcriptional responses of three model diatoms to nitrate limitation of growth“. In: *Frontiers in Marine Science* 1.March, pp. 1–15 (cit. on pp. 92, 149–154, 162, 163).

Bender, S. J., M. S. Parker, and E. V. Armbrust (2012). „Coupled Effects of Light and Nitrogen Source on the Urea Cycle and Nitrogen Metabolism over a Diel Cycle in the Marine Diatom Thalassiosira pseudonana“. In: *Protist* 163.2, pp. 232–251 (cit. on pp. 26, 92, 152).

Benjamini, Y. and Y. Hochberg (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. arXiv: 95/57289 [0035-9246] (cit. on pp. 157, 158, 200).

Berges, J. A. and P. J. Harrison (1995). „Nitrate reductase activity quantitatively predicts the rate of nitrate incorporation under steady state light limitation: A revised assay and characterization of the enzyme in three species of marine phytoplankton“. In: *Limnology and Oceanography* (cit. on p. 91).

Bernardes, J. S., F. R. Vieira, G. Zaverucha, and A. Carbone (2015). „A multi-objective optimization approach accurately resolves protein domain architectures“. In: *Bioinformatics* 32.3, pp. 345–353 (cit. on p. 94).

Berthon, V., A. Bouchez, and F. Rimet (2011). „Using diatom life-forms and ecological guilds to assess organic pollution and trophic level in rivers: A case study of rivers in south-eastern France“. In: *Hydrobiologia* (cit. on p. 90).

Bhadury, P., B. Song, and B. B. Ward (2011). „Intron features of key functional genes mediating nitrogen metabolism in marine phytoplankton“. In: *Marine Genomics* 4.3, pp. 207–213 (cit. on pp. 149, 152).

Bhattacharya, D., D. C. Price, H. S. Yoon, et al. (2012). „Single cell genome analysis supports a link between phagotrophy and primary plastid endosymbiosis“. In: *Scientific Reports* (cit. on p. 232).

Biswas, S. B., M. Mcdonald, D. S. Lundberg, J. L. Dangl, and V. Jojic (2015). „Research in Computational Molecular Biology“. In: 9029, pp. 32–33 (cit. on p. 33).

Bonachela, J. A., C. A. Klausmeier, K. F. Edwards, E. Litchman, and S. A. Levin (2016). „The role of phytoplankton diversity in the emergent oceanic stoichiometry“. In: *Journal of Plankton Research* 38.4, pp. 1021–1035 (cit. on p. 1).

Bonan, G. B. and S. C. Doney (2018). „Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models“. In: *Science* 359.6375 (cit. on p. 195).

Bork, P., C. Bowler, C. De Vargas, et al. (2015). *Tara Oceans studies plankton at Planetary scale* (cit. on p. 33).

Bottin, M., J. Soininen, D. Alard, and J. Rosebery (2016). „Diatom cooccurrence shows less segregation than predicted from niche modeling“. In: *PLoS ONE* 11.4, pp. 1–18 (cit. on p. 18).

Bouman, H. A., T. Platt, S. Sathyendranath, et al. (2003). „Temperature as indicator of optical properties and community structure of marine phytoplankton: Implications for remote sensing“. In: *Marine Ecology Progress Series* (cit. on p. 219).

Bowler, C., A. Vardi, and A. E. Allen (2010). „Oceanographic and biogeochemical insights from diatom genomes“. In: *Annual Review of Marine Science* (cit. on pp. 21, 27).

Boyd, C. M. and D. Gradmann (1999). „Electrophysiology of the marine diatom Coscinodiscus wailesii: III. Uptake of nitrate and ammonium“. In: *Journal of Experimental Botany* (cit. on p. 26).

Boyer, T. P., J. I. Antonov, O. K. Baranova, et al. (2013). „World Ocean Database 2013“. In: *Sydney Levitus, Ed.; Alexey Mishonoc, Technical Ed.* arXiv: V5NZ85MT [10.7289] (cit. on p. 130).

Bracco, A., A. Provenzale, and I. Scheuring (2000). „Mesoscale vortices and the paradox of the plankton“. In: *Proceedings of the Royal Society B: Biological Sciences* 267.1454, pp. 1795–1800. arXiv: 2082 (cit. on p. 19).

Brown, K. L., K. I. Twing, and D. L. Robertson (2009). „Unraveling the regulation of nitrogen assimilation in the marine diatom thalassiosira pseudonana (bacillariophyceae): Diurnal variations in transcript levels for five genes involved in nitrogen assimilation“. In: *Journal of Phycology* (cit. on p. 91).

Brown, S. P., A. M. Veach, A. R. Rigdon-Huss, et al. (2015). „Scraping the bottom of the barrel: Are rare high throughput sequences artifacts?“ In: *Fungal Ecology* 13, pp. 221–225 (cit. on pp. 68, 83).

Browning, T. J., E. P. Achterberg, I. Rapp, et al. (2017). „Nutrient co-limitation at the boundary of an oceanic gyre“. In: *Nature* 551.7679, pp. 242–246 (cit. on p. 195).

Brzezinski, M. A., J. W. Krause, M. J. Church, et al. (2011). „The annual silica cycle of the North Pacific subtropical gyre“. In: *Deep-Sea Research Part I: Oceanographic Research Papers* 58.10, pp. 988–1001 (cit. on p. 20).

Buitenhuis, E. T., M. Vogt, R. Moriarty, et al. (2013). „MAREDAT: Towards a world atlas of MARine Ecosystem DATa“. In: *Earth System Science Data* 5.2, pp. 227–239 (cit. on pp. 17, 219).

Buttigieg, P. L., E. Fadeev, C. Bienhold, et al. (2018). „Marine microbes in 4D — using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health“. In: *Current Opinion in Microbiology* 43, pp. 169–185 (cit. on p. 231).

Cadotte, M. W., K. Carscadden, and N. Mirotchnick (2011). „Beyond species: Functional diversity and the maintenance of ecological processes and services“. In: *Journal of Applied Ecology* 48.5, pp. 1079–1087 (cit. on p. 9).

Caputi, L., Q. Carradec, D. Eveillard, et al. (Submitted). „Community-level responses to iron availability in open ocean planktonic ecosystems“. In: (cit. on p. 80).

Caron, D. A. and P. D. Countway (2009). „Hypotheses on the role of the protistan rare biosphere in a changing world“. In: *Aquatic Microbial Ecology* (cit. on pp. 43, 70).

Carradec, Q., E. Pelletier, C. Da Silva, et al. (2018). „A global ocean atlas of eukaryotic genes“. In: *Nature Communications* 9.373, 1:13 (cit. on pp. 36, 94, 97, 127, 156).

Casteleyn, G., F. Leliaert, T. Backeljau, et al. (2010). „Limits to gene flow in a cosmopolitan marine planktonic diatom“. In: *Proceedings of the National Academy of Sciences* 107.29, pp. 12952–12957 (cit. on p. 19).

Casteleyn, G., N. G. Adams, P. Vanormelingen, et al. (2009). „Natural hybrids in the marine diatom Pseudo-nitzschia pungens (Bacillariophyceae): Genetic and morphological evidence“. In: *Protist* (cit. on p. 233).

Cermeno, P. and P. G. Falkowski (2009). „Controls on diatom biogeography in the ocean“. In: *Science* 325.5947, pp. 1539–1541 (cit. on p. 18).

Cermeño, P., I. G. Teixeira, M. Branco, F. G. Figueiras, and E. Marañón (2014). „Sampling the limits of species richness in marine phytoplankton communities“. In: *Journal of Plankton Research* 36.4, pp. 1135–1139 (cit. on pp. 43, 44).

Chaffron, S., H. Rehrauer, J. Pernthaler, and C. Von Mering (2010). „A global network of coexisting microbes from environmental and whole-genome sequence data“. In: *Genome Research* (cit. on p. 33).

Chapin, F. S., E. S. Zavaleta, V. T. Eviner, et al. (2000). „Consequences of changing biodiversity.“ In: *Nature* 405.6783, pp. 234–42. arXiv: arXiv:1011.1669v3 (cit. on p. 8).

Chaudhary, C., H. Saeedi, and M. J. Costello (2016). „Bimodality of latitudinal gradients in marine species richness“. In: *Trends in Ecology and Evolution* 31.9, pp. 670–676 (cit. on p. 47).

Chepurnov, V., D. Mann, K. Sabbe, and W. Vyverman (2004). „Experimental studies on sexual reproduction in diatoms“. In: *Int. Rev. Cytol* 237, pp. 91–154 (cit. on p. 3).

Chesson, P. (2000). „Mechanisms of maintenance of species diversity“. In: *Annual Review of Ecology and Systematics* 31.1, pp. 343–366. arXiv: annurev.ecolsys.31.1.343 [10.1146] (cit. on p. 11).

Chust, G., X. Irigoien, J. Chave, and R. P. Harris (2013). „Latitudinal phytoplankton distribution and the neutral theory of biodiversity“. In: *Global Ecology and Biogeography* 22.5, pp. 531–543 (cit. on pp. 13, 19, 47, 125, 144).

Clarke, K. R., P. J. Somerfield, and M. G. Chapman (2006). „On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages“. In: *Journal of Experimental Marine Biology and Ecology* 330.1, pp. 55–80 (cit. on pp. 128, 158).

Clayton, S., S. Dutkiewicz, O. Jahn, and M. J. Follows (2013). „Dispersal, eddies, and the diversity of marine phytoplankton“. In: *Limnology and Oceanography: Fluids and Environments* (cit. on pp. 18, 19, 46).

Coale, K. H., K. S. Johnson, S. E. Fitzwater, et al. (1996). „A massive phytoplankton bloom induced by an ecosystem-scale iron fertilization experiment in the equatorial Pacific Ocean“. In: *Nature* (cit. on p. 86).

Cohen, N. R., K. A. Ellis, R. H. Lampe, et al. (2017). „Diatom transcriptional and physiological responses to changes in iron bioavailability across ocean provinces“. In: *Frontiers in Marine Science* 4.November, pp. 1–20 (cit. on pp. 144, 150).

Coles, V. J., M. R. Stukel, M. T. Brooks, et al. (2017). „Ocean biogeochemistry modeled with emergent trait-based genomics“. In: *Science* 358.6367, pp. 1149–1154 (cit. on pp. 16, 124, 195, 196).

Colios, Y. (1982). „Transient situations in nitrate assimilation by marine diatoms. Changes in nitrate and nitrite following a nitrate perturbation“. In: *Limnology and Oceanography* (cit. on p. 27).

Comtet, T., A. Sandionigi, F. Viard, and M. Casiraghi (2015). „DNA (meta)barcoding of biological invasions: a powerful tool to elucidate invasion processes and help managing aliens“. In: *Biological Invasions* (cit. on p. 44).

Conant, G. C., J. A. Birchler, and J. C. Pires (2014). *Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time* (cit. on p. 119).

Conley, D. J. and J. C. Carey (2015). „Biogeochemistry: Silica cycling over geologic time“. In: *Nature Geoscience* 8.6, pp. 431–432 (cit. on p. 22).

Conway, H. L., P. J. Harrison, and C. O. Davis (1976). „Marine diatoms grown in chemostats under silicate or ammonium limitation. II. Transient response of Skeletonema costatum to a single addition of the limiting nutrient“. In: *Marine Biology* (cit. on p. 25).

Cornwell, W. K. and D. D. Ackerly (2009). „Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California“. In: *Ecological Monographs* (cit. on p. 124).

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner (2004). „WebLogo: A sequence logo generator“. In: *Genome Research* 14.6, pp. 1188–1190 (cit. on p. 96).

Dagenais Bellefeuille, S. and D. Morse (2016). „The main nitrate transporter of the dinoflagellate Lingulodinium polyedrum is constitutively expressed and not responsible for daily variations in nitrate uptake rates“. In: *Harmful Algae* 55, pp. 272–281 (cit. on p. 104).

Dakos, V., E. Benincà, E. H. Van Nes, et al. (2009). „Interannual variability in species composition explained as seasonally entrained chaos“. In: *Proceedings of the Royal Society B: Biological Sciences* (cit. on p. 10).

D'Alelio, D., D. Eveillard, V. Coles, et al. (Submitted). „Modelling the complexity of plankton communities exploiting omics potential: from present challenges to a feasible pipeline“. In: (cit. on p. 16).

De Benedictis, P. (1973). „On the correlations between certain diversity indices“. In: *American Naturalist* 107, pp. 295–302 (cit. on p. 8).

De Monte, S., A. Soccodato, S. Alvain, and F. D'Ovidio (2013). „Can we detect oceanic biodiversity hotspots from space?“ In: *ISME Journal* 7.10, pp. 2054–2056 (cit. on pp. 125, 144).

De Vargas, C., S. Audic, N. Henry, et al. (2015). „Eukaryotic plankton diversity in the sunlit ocean“. In: *Science* 348.6237. arXiv: `9809069v1 [arXiv:gr-qc]` (cit. on pp. 48, 49).

De'ath, G. (2007). „Boosted regression trees for ecological modeling and prediction“. In: *Ecology* 88.1, pp. 243–251 (cit. on p. 55).

Dell'aquila, G., M. I. Ferrante, M. Gherardi, et al. (2017). „Nutrient consumption and chain tuning in diatoms exposed to storm-like turbulence“. In: *Scientific Reports* 7.1, pp. 1–11 (cit. on pp. 5, 150).

Denman, K. L. (2008). *Climate change, ocean processes and ocean iron fertilization* (cit. on p. 22).

Dentener, F. J. (2006). *Global maps of atmospheric nitrogen deposition, 1860, 1993, and 2050* (cit. on p. 24).

Dodds, W., J. Jones, and E. Welch (1998). *Suggested-classification of stream trophic state: distributions of temperate stream types by chlorophyll, total nitrogen and total phosphorus* (cit. on p. 130).

Doney, S. C., M. Ruckelshaus, J. Emmett Duffy, et al. (2012). „Climate change impacts on marine ecosystems“. In: *Annual Review of Marine Science* 4.1, pp. 11–37 (cit. on p. 1).

Dortch, Q. (1990). „The interaction between ammonium and nitrate uptake in phytoplankton“. In: *Marine Ecology Progress Series* 61, pp. 183–201 (cit. on pp. 23, 25).

Dortch, Q., P. A. Thompson, and P. J. Harrison (1991). „Short-term interaction between nitrate and ammonium uptake in Thalassiosira pseudonana: Effect of preconditioning nitrogen source and growth rate“. In: *Marine Biology* (cit. on p. 27).

D'Ovidio, F., S. De Monte, S. Alvain, Y. Dandonneau, and M. Levy (2010). „Fluid dynamical niches of phytoplankton types“. In: *Proceedings of the National Academy of Sciences* 107.43, pp. 18366–18370 (cit. on pp. 46, 56).

Duce, R. A., J. LaRoche, K. Altieri, et al. (2008). *Impacts of atmospheric anthropogenic nitrogen on the open ocean* (cit. on p. 24).

Dunthorn, M., J. Otto, S. A. Berger, et al. (2014). „Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context“. In: *Molecular Biology and Evolution* (cit. on p. 45).

Durkin, C. A., A. Marchetti, S. J. Bender, et al. (2012). „Frustule-related gene transcription and the influence of diatom community composition on silica precipitation in an iron-limited environment“. In: *Limnology and Oceanography* (cit. on p. 22).

Dutkiewicz, S., M. J. Follows, and J. G. Bragg (2009). „Modeling the coupling of ocean ecology and biogeochemistry“. In: *Global Biogeochemical Cycles* 23.4, pp. 1–15 (cit. on pp. 12, 48, 79, 86, 198).

Dutkiewicz, S., A. E. Hickman, O. Jahn, et al. (2015). „Capturing optically important constituents and properties in a marine biogeochemical and ecosystem model“. In: *Biogeosciences* 12.14, pp. 4447–4481 (cit. on pp. 5, 15, 197, 198, 220, 221).

Dutkiewicz, S., B. A. Ward, F. Monteiro, and M. J. Follows (2012). „Interconnection of nitrogen fixers and iron in the Pacific Ocean: Theory and numerical simulations“. In: *Global Biogeochemical Cycles* (cit. on p. 198).

Dyhrman, S. T., B. D. Jenkins, T. A. Rynearson, et al. (2012). „The transcriptome and proteome of the diatom Thalassiosira pseudonana reveal a diverse phosphorus stress response“. In: *PLoS ONE* 7.3 (cit. on p. 152).

Edgar, R. C. (2004). „MUSCLE: Multiple sequence alignment with high accuracy and high throughput“. In: *Nucleic Acids Research* 32.5, pp. 1792–1797. arXiv: NIHMS150003 (cit. on p. 95).

Edwards, K. F., E. Litchman, and C. A. Klausmeier (2013a). „Functional traits explain phytoplankton community structure and seasonal dynamics in a marine ecosystem“. In: *Ecology Letters* 16.1, pp. 56–63 (cit. on pp. 125, 141).

– (2013b). „Functional traits explain phytoplankton responses to environmental gradients across lakes of the United States“. In: *Ecology* (cit. on p. 141).

Edwards, K. F., M. K. Thomas, C. A. Klausmeier, and E. Litchman (2012). „Allometric scaling and taxonomic variation in nutrient utilization traits and maximum growth rate of phytoplankton“. In: *Limnology and Oceanography* 57.2, pp. 554–566 (cit. on p. 201).

– (2015). „Light and growth in marine phytoplankton: Allometric, taxonomic, and environmental variation“. In: *Limnology and Oceanography* 60.2, pp. 540–552. arXiv: arXiv:1011.1669v3 (cit. on pp. 15, 21).

Eilertsen, H. C., S. Sandberg, and H. Tollefsen (1995). „Photoperiodic control of diatom spore growth: A theory to explain the onset of phytoplankton blooms“. In: *Marine Ecology Progress Series* (cit. on p. 4).

Elith, J., J. R. Leathwick, and T. Hastie (2008). „A working guide to boosted regression trees“. In: *Journal of Animal Ecology* 77.4, pp. 802–813 (cit. on pp. 54, 159, 181).

Eppley, R. W., J. N. Rogers, and J. J. Mccarthy (1969). „Half-saturation constants for uptake of nitrate and ammonium by marine phytoplankton“. In: *Methods* (cit. on pp. 201, 215).

Evans, K. M., A. H. Wortley, and D. G. Mann (2007). „An Assessment of Potential Diatom "Barcode" Genes (cox1, rbcL, 18S and ITS rDNA) and their Effectiveness in Determining Relationships in Sellaphora (Bacillariophyta)“. In: *Protist* (cit. on p. 44).

Evans, K. M., A. H. Wortley, G. E. Simpson, V. A. Chepurnov, and D. G. Mann (2008). „A molecular systematic approach to explore diversity within the Sellaphora pupula species complex (Bacillariophyta)“. In: *Journal of Phycology* (cit. on p. 44).

Falkowski, P. G., R. T. Barber, and V. Smetacek (1998). *Biogeochemical controls and feedbacks on ocean primary production* (cit. on p. 219).

Falkowski, P. G., M. E. Katz, A. H. Knoll, et al. (2004). „The evolution of modern eukaryotic phytoplankton“. In: *Science* 305.5682, pp. 354–360 (cit. on p. 1).

Falkowski, P. G. and M. J. Oliver (2007). „Mix and match: How climate selects phytoplankton“. In: *Nature Reviews Microbiology* (cit. on p. 171).

Falony, G., S. Vieira-Silva, and J. Raes (2015). „Microbiology meets big data: The case of gut microbiota–derived trimethylamine“. In: *Annual Review of Microbiology* (cit. on p. 28).

Fenchel, T. and B. J. Finlay (2004). „The ubiquity of small species: patterns of local and global diversity“. In: *BioScience* 54.8, p. 777 (cit. on p. 18).

Ficetola, G. F., J. Pansu, A. Bonin, et al. (2015). „Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data“. In: *Molecular Ecology Resources* 15.3, pp. 543–556 (cit. on p. 45).

Ficetola, G. F., P. Taberlet, and E. Coissac (2016). „How to limit false positives in environmental DNA and metabarcoding?“ In: *Molecular Ecology Resources* 16.3, pp. 604–607 (cit. on p. 45).

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski (1998). „Primary production of the biosphere: Integrating terrestrial and oceanic components“. In: *Science* 281.5374, pp. 237–240. arXiv: 1011.1669 (cit. on p. 22).

Finkel, Z. V., J. Beardall, K. J. Flynn, et al. (2010). *Phytoplankton in a changing world: Cell size and elemental stoichiometry* (cit. on p. 89).

Finlay, B. J. (2002). „Global dispersal of free-living microbial eukaryote species.“ In: *Science (New York, N.Y.)* 296.5570, pp. 1061–3 (cit. on p. 18).

Finlay, B. J., E. B. Monaghan, and S. C. Maberly (2002). „Hypothesis: the rate and scale of dispersal of freshwater diatom species is a function of their global abundance.“ In: *Protist* 153.3, pp. 261–273 (cit. on p. 18).

Flori, S., P. H. Jouneau, B. Bailleul, et al. (2017). „Plastid thylakoid architecture optimizes photosynthesis in diatoms“. In: *Nature Communications* (cit. on p. 21).

Foissner, W. (2006). „Biogeography and dispersal of micro-organisms: A review emphasizing protists“. In: *Acta Protozoologica* 45.2, pp. 111–136 (cit. on p. 18).

Follows, M. J., S. Dutkiewicz, S. Grant, and S. W. Chisholm (2007). „Emergent biogeography of microbial communities in a model ocean“. In: *Science* 315.5820, pp. 1843–1846 (cit. on pp. 18, 198, 220, 221, 235).

Follows, M. J. and S. Dutkiewicz (2011). „Modeling diverse communities of marine microbes“. In: *Annual Review of Marine Science* 3.1, pp. 427–451 (cit. on p. 194).

Force, A., M. Lynch, F. B. Pickett, et al. (1999). „Preservation of duplicate genes by complementary, degenerative mutations“. In: *Genetics* 151.4, pp. 1531–1545 (cit. on p. 155).

Fraisier, V., A. Gojon, P. Tillard, and F. Daniel-Vedele (2000). „Constitutive expression of a putative high-affinity nitrate transporter in Nicotiana plumbaginifolia: Evidence for post-transcriptional regulation by a reduced nitrogen source“. In: *Plant Journal* (cit. on p. 155).

Freeman, E. A. and G. G. Moisen (2008). „A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa“. In: *Ecological Modelling* (cit. on p. 161).

Frölicher, T. L., K. B. Rodgers, C. A. Stock, and W. W. L. Cheung (2016). „Sources of uncertainties in 21st century projections of potential ocean ecosystem stressors“. In: *Global Biogeochemical Cycles*, pp. 1224–1243 (cit. on p. 195).

Frommer, W. B., M. Kwart, B. Hirner, et al. (1994). „Transporters for nitrogenous compounds in plants“. In: *Plant Molecular Biology* (cit. on p. 27).

Fryxell, G. and M. Villac (1999). „Toxic and harmful marine diatoms“. In: *The Diatoms: Applications for the Environmental and Earth Sciences*, pp. 419–428 (cit. on p. 5).

Fung, I. Y., S. K. Meyn, I. Tegen, et al. (2000). „Iron supply and demand in the upper ocean“. In: *Global Biogeochemical Cycles* (cit. on p. 182).

Galhardo, R. S., P. J. Hastings, and S. M. Rosenberg (2007). *Mutation as a stress response and the regulation of evolvability*. Vol. 42. 5, pp. 399–435 (cit. on p. 84).

Galván, A. and E. Fernández (2001). *Eukaryotic nitrate and nitrite transporters* (cit. on p. 27).

Galván, A., J. Rexach, V. Mariscal, and E. Fernández (2002). „Nitrite transport to the chloroplast in Chlamydomonas reinhardtii: Molecular evidence for a regulated process“. In: *Journal of Experimental Botany* (cit. on p. 27).

Gao, Y., G. J. Smith, and R. S. Alberte (1993). „Nitrate reductase from the marine diatom Skeletonema costatum (biochemical and immunological characterization).“ In: *Plant Physiol* (cit. on p. 91).

Gause, G. F. (1934). „Experimental analysis of Vito Volterra's mathematical theory of the struggle for existence“. In: *Science*. arXiv: `arXiv:1011.1669v3` (cit. on p. 10).

Gersonde, R. and D. Harwood (1990). „Lower cretaceous diatoms from ODP leg 113 site 693 (Weddell Sea). Part 1: Vegetative cells“. In: *Proceedings of the Ocean Drilling Program, 113 Scientific Reports* 113 (cit. on p. 2).

Gilbert, J. A., J. K. Jansson, and R. Knight (2014). *The Earth Microbiome project: Successes and aspirations* (cit. on p. 33).

Gilbert, J. A., R. O'Dor, N. King, and T. M. Vogel (2011). „The importance of metagenomic surveys to microbial ecology: or why Darwin would have been a metagenomic scientist“. In: *Microbial Informatics and Experimentation* 1.1, p. 5 (cit. on p. 33).

Gill, S. R., M. Pop, R. T. DeBoy, et al. (2006). „Metagenomic analysis of the human distal gut microbiome“. In: *Science* (cit. on pp. 28, 30).

Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field (1990). „Genetic diversity in Sargasso Sea bacterioplankton“. In: *Nature*. arXiv: NIHMS150003 (cit. on p. 31).

Gitav, H. and I. R. Noble (1997). „What are functional types and how should we seek them?“ In: *Plant functional types: their relevance to ecosytem properties and global change*. Ed. by T. M. Smith, H. Shugart, and F. Woodward. Cambridge: Cambridge university press, pp. 3–19 (cit. on p. 9).

Glibert, P. M. and G. M. Berg (2009). „Nitrogen form, fate and phytoplankton composition“. In: *Experimental Ecosystems and Scale: Tools for Understanding and Managing Coastal Ecosystems*, pp. 183–189 (cit. on p. 25).

Glibert, P. M., F. P. Wilkerson, R. C. Dugdale, et al. (2016). „Pluses and minuses of ammonium and nitrate uptake and assimilation by phytoplankton and implications for productivity and community composition, with emphasis on nitrogen-enriched conditions“. In: *Limnology and Oceanography* 61.1, pp. 165–197 (cit. on pp. 25, 26, 29, 92).

Gonzalez, J. M., M. C. Portillo, P. Belda-Ferre, and A. Mira (2012). „Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities“. In: *PLoS ONE* (cit. on p. 82).

Gotelli, N. J. and R. K. Colwell (2001). „Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness“. In: *Ecology Letters* 4.4, pp. 379–391. arXiv: 2675 (cit. on pp. 44, 98).

Gravel, D., C. D. Canham, M. Beaudet, and C. Messier (2006). „Reconciling niche and neutrality: The continuum hypothesis“. In: *Ecology Letters* 9.4, pp. 399–409. arXiv: 2651 (cit. on p. 13).

Gray, J. S. (2001). „Marine diversity : the paradigms in patterns of species richness examined“. In: *Scientia Marina* 65.Suppl. 2, pp. 41–56 (cit. on p. 48).

Green, J. L., B. J. Bohannan, and R. J. Whitaker (2008). „Microbial biogeography : From taxonomy to traits“. In: *Science* 320.5879, pp. 1039–1043 (cit. on p. 124).

Grigoriev, I. V., R. Nikitin, S. Haridas, et al. (2014). „MycoCosm portal: Gearing up for 1000 fungal genomes“. In: *Nucleic Acids Research* (cit. on p. 31).

Groendahl, S., M. Kahlert, and P. Fink (2017). „The best of both worlds: A combined approach for analyzing microalgal diversity via metabarcoding and morphology-based methods“. In: *PLoS ONE* 12.2, pp. 1–15 (cit. on p. 45).

Grover, J. P. (1991). „Resource competition in a variable environment: Phytoplankton growing according to the variable-internal-stores model“. In: *The American Naturalist* 138.4, pp. 811–835 (cit. on p. 5).

Gu, Z., D. Nicolae, H. H. S. Lu, and W. H. Li (2002). *Rapid divergence in expression between duplicate genes inferred from microarray data* (cit. on p. 155).

Guidi, L., L. Legendre, G. Reygondeau, et al. (2015). „A new look at ocean carbon remineralization for estimating deepwater sequestration“. In: *Global Biogeochemical Cycles* (cit. on p. 175).

Guiry, M. D. (2012). *How many species of algae are there?* arXiv: NIHMS150003 (cit. on pp. 106, 231).

Hamm, C. E., R. Merkel, O. Springer, et al. (2003). „Architecture and material properties of diatom shells provide effective mechanical protection“. In: *Nature* 421.6925, pp. 841–843 (cit. on pp. 3, 12).

Hamsher, S. E., K. M. Evans, D. G. Mann, A. Poulíčková, and G. W. Saunders (2011). „Barcoding diatoms: Exploring alternatives to COI-5P“. In: *Protist* (cit. on p. 44).

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman (1998). „Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products“. In: *Chemistry and Biology* (cit. on p. 28).

Hansen, J., M. Sato, R. Ruedy, et al. (2006). „Global temperature change.“ In: *Proceedings of the National Academy of Sciences of the United States of America* (cit. on p. 182).

Hays, G. C., A. J. Richardson, and C. Robinson (2005). „Climate change and marine plankton“. In: *Trends in Ecology and Evolution* 20.6 SPEC. ISS. Pp. 337–344. arXiv: arXiv:1011.1669v3 (cit. on p. 1).

Hellweger, F. L., E. Van Sebille, and N. D. Fredrick (2014). „Biogeographic patterns in ocean microbes emerge in a neutral agent-based model“. In: *Science* (cit. on p. 18).

Hendry, K. R., A. O. Marron, F. Vincent, et al. (2018). „Competition between silicifiers and non-silicifiers in the past and present ocean and its evolutionary impacts“. In: *Frontiers in Marine Science* 5.February, pp. 1–21 (cit. on p. 36).

Hess, M., A. Sczyrba, R. Egan, et al. (2011). „Metagenomic discovery of biomass-degrading genes and genomes from cow rumen“. In: *Science*. arXiv: arXiv:1011.1669v3 (cit. on p. 30).

Hickman, A. E., S. Dutkiewicz, R. G. Williams, and M. J. Follows (2010). „Modelling the effects of chromatic adaptation on phytoplankton community structure in the oligotrophic ocean“. In: *Marine Ecology Progress Series* (cit. on p. 198).

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith (2017). „dismo: Species Distribution Modeling. R package version 1.1-4.“ In: (cit. on pp. 55, 159).

Hildebrand, M. (2005). „Cloning and functional characterization of ammonium transporters from the marine diatom Cylindrotheca fusiformis (Bacillariophyceae)“. In: *J. Phycol.* 113, pp. 105–113 (cit. on pp. 26, 149, 152).

Hildebrand, M. and K. Dahlin (2000). „Nitrate transporter genes from the diatom Cylindrotheca fusiformis (Bacillariophyceae): mRNA levels controlled by nitrogen source and by the cell cycle“. In: *J. Phycol.* 713.April, pp. 702–713 (cit. on p. 150).

Hill, T. C. J., K. A. Walsh, and J. A. Harris (2003). „Using ecological diversity measures with bacterial communities“. In: *FEMS Microbiology Ecology* 43.April, pp. 1–11 (cit. on p. 8).

Hillebrand, H. (2004). „On the generality of the latitudinal diversity gradient“. In: *American Naturalist* 163.2, pp. 192–211. arXiv: arXiv:1011.1669v3 (cit. on p. 47).

Hilton, J. A., R. A. Foster, H. James Tripp, et al. (2012). „Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont“. In: *Nature Communications* (cit. on p. 26).

Hirai, J., M. Kuriyama, T. Ichikawa, K. Hidaka, and A. Tsuda (2015). „A metagenetic approach for revealing community structure of marine planktonic copepods“. In: *Molecular Ecology Resources* (cit. on p. 82).

Hiraoka, S., C.-c. Yang, and W. Iwasaki (2016). „Metagenomics and bioinformatics in microbial ecology: Current status and beyond“. In: *Microbes and environments* 31.3, pp. 204–212 (cit. on pp. 30, 32).

Hockin, N. L., T. Mock, F. Mulholland, S. Kopriva, and G. Malin (2012). „The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants“. In: *Plant Physiology* (cit. on pp. 27, 28).

Hood, R. R., E. A. Laws, R. A. Armstrong, et al. (2006). „Pelagic functional group modeling: Progress, challenges and prospects“. In: *Deep-Sea Research Part II: Topical Studies in Oceanography* 53.5-7, pp. 459–512 (cit. on pp. 8, 89).

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography*. arXiv: arXiv:1011.1669v3 (cit. on p. 13).

Huisman, J. and F. J. Weissing (2000). „Biodiversity - Coexistence and resource competition - Reply“. In: *Nature* 407.6805, p. 694 (cit. on p. 11).

Huisman, J., A. M. Johansson, E. O. Folmer, and F. J. Weissing (2001). „Towards a solution of the plankton paradox: The importance of physiology and life history“. In: *Ecology Letters* 4.5, pp. 408–411 (cit. on p. 10).

Huisman, J. and F. J. Weissing (1999). „Biodiversity of plankton by species oscillations and chaos“. In: *Nature* 402.6760, pp. 407–410 (cit. on pp. 10, 11).

Hurlbert, S. H. (1971). „The nonconcept of species diversity: A critique and alternative parameters“. In: *Ecology* (cit. on p. 7).

Hutchinson, G. E. (1961). „The paradox of the plankton“. In: *The American Naturalist* 95.882, pp. 137–145. arXiv: 96/4702-0006{\$}02 [0003-0147] (cit. on pp. 10, 12, 48).

Irigoien, X., J. Huisman, and R. P. Harris (2004). „Global biodiversity patterns of marine phytoplankton and zooplankton“. In: *Nature* 429.6994, pp. 863–867 (cit. on pp. 47, 66).

Irwin, A. J., Z. V. Finkel, O. M. Schofield, and P. G. Falkowski (2006). „Scaling-up from nutrient physiology to the size-structure of phytoplankton communities“. In: *Journal of Plankton Research* (cit. on pp. 89, 221).

Iudicone, D., A. Amato, M. I. Ferrante, and M. Ribera d'Alcalà (2016). „Diatoms morphology and gene expression in turbulence“. In: *American Geophysical Union, Ocean Sciences Meeting* (cit. on p. 5).

Jeong, J. (2004). „A nodule-specific dicarboxylate transporter from alder is a member of the peptide transporter family“. In: *Plant Physiology* (cit. on p. 27).

Johnson, Z. I., E. R. Zinser, A. Coe, et al. (2006). „Partitioning among Prochlorococcus ecotypes along environmental gradients“. In: *Science* 311.March, pp. 1737–1740 (cit. on p. 125).

Jones, D. T., W. R. Taylor, and J. M. Thornton (1992). „The rapid generation of mutation data matrices from protein sequences“. In: *Bioinformatics* 8.3, pp. 275–282 (cit. on p. 96).

Kaiser, D., N. Kowalski, and J. J. Waniek (2017). „Effects of biofouling on the sinking behavior of microplastics“. In: *Environmental Research Letters* 12.12 (cit. on p. 6).

Kang, L.-k. and J. Chang (2014). „Sequence diversity of ammonium transporter genes in clutured and natural species of marine phytoplankton“. In: *Jounal of Marine Science and Technology* (cit. on p. 25).

Kang, L.-K., G.-C. Gong, Y.-H. Wu, and J. Chang (2015). „The expression of nitrate transporter genes reveals different nitrogen statuses of dominant diatom groups in the southern East China Sea“. In: *Molecular Ecology* 24, pp. 1374–1386 (cit. on pp. 149, 154).

Kanno, Y., A. Hanada, Y. Chiba, et al. (2012). „Identification of an abscisic acid transporter by functional screening using the receptor complex as a sensor“. In: *Proceedings of the National Academy of Sciences* (cit. on p. 27).

Karsenti, E., S. G. Acinas, P. Bork, et al. (2011). „A holistic approach to marine Eco-systems biology“. In: *PLoS Biology* (cit. on pp. 9, 35).

Kattge, J., S. Díaz, S. Lavorel, et al. (2011). „TRY - a global database of plant traits“. In: *Global Change Biology* 17.9, pp. 2905–2935 (cit. on p. 15).

Katz, M. E., K. Fennel, and P. G. Falkowski (2007). „Geochemical and biological consequences of phytoplakton evolution“. In: chap. 18, pp. 405–430 (cit. on p. 1).

Keeling, P. J., F. Burki, H. M. Wilcox, et al. (2014). „The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing“. In: *PLoS Biology* 12.6. arXiv: `0402594v3 [arXiv:cond-mat]` (cit. on pp. 95, 96, 232).

Kelly, M., N. Boonham, S. Juggins, et al. (2018). *A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers* (cit. on p. 44).

Kembel, S. W., M. Wu, J. A. Eisen, and J. L. Green (2012). „Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance“. In: *PLoS Computational Biology* (cit. on p. 82).

Kemp, A. E. and T. A. Villareal (2018). „The case of the diatoms and the muddled mandalas: Time to recognize diatom adaptations to stratified waters“. In: *Progress in Oceanography* 167.April, pp. 138–149 (cit. on p. 231).

Kermarrec, L., A. Franc, F. Rimet, et al. (2013). „Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: A test for freshwater diatoms“. In: *Molecular Ecology Resources* (cit. on p. 44).

Kermarrec, L., A. Franc, F. Rimet, and P. Chaumeil (2014). „A next-generation sequencing approach to river biomonitoring using benthic diatoms“. In: *Molecular Approaches in Freshwater Ecology* 13.August 2013, pp. 349–363 (cit. on p. 44).

Kishi, M. J., M. Kashiwai, D. M. Ware, et al. (2007). „NEMURO-a lower trophic level model for the North Pacific marine ecosystem“. In: *Ecological Modelling* (cit. on p. 219).

Kohonen, T. (2001). „Self-Organizing Maps“. In: *Springer Series in Information Sciences* 30, p. 501. arXiv: `arXiv:1011.1669v3` (cit. on p. 59).

Kókai, Z., I. Bácsi, P. Török, et al. (2015). „Halophilic diatom taxa are sensitive indicators of even short term changes in lowland lotic systems“. In: *Acta Botanica Croatica* (cit. on p. 90).

Kolde, R. (2015). „pheatmap : Pretty Heatmaps“. In: *R package version 1.0.8* (cit. on pp. 199, 201).

Kooistra, W. H., R. Gersonde, L. K. Medlin, and D. G. Mann (2007). „The origin and evolution of the diatoms. Their adaptation to a planktonic existence.“ In: *Evolution of primary producers in the sea* (cit. on p. 6).

Kooistra, W. H. and L. K. Medlin (1996). „Evolution of the diatoms (Bacillariophyta) IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record“. In: *Molecular Phylogenetics and Evolution* (cit. on p. 2).

Kotur, Z., S. E. Unkles, and A. D. M. Glass (2016). *Comparisons of the Arabidopsis thaliana high-affinity Nitrate Transporter complex AtNRT2.1/AtNAR2.1 and the Aspergillus nidulans AnNRTA: structure function considerations* (cit. on p. 104).

Kozubowski, T. and K. Podgórski (2012). *Laplace probability distributions and related stochastic processes* (cit. on p. 56).

Kraft, N. J. B., P. B. Adler, O. Godoy, et al. (2015). „Community assembly, coexistence and the environmental filtering metaphor“. In: *Functional Ecology* 29.5, pp. 592–599 (cit. on p. 19).

Krause, J. W., M. A. Brzezinski, S. B. Baines, et al. (2017). „Picoplankton contribution to biogenic silica stocks and production rates in the Sargasso Sea“. In: *Global Biogeochemical Cycles* 31.5, pp. 762–774 (cit. on p. 22).

Krohn-Molt, I., M. Alawi, K. U. Förstner, et al. (2017). „Insights into microalga and bacteria interactions of selected phycosphere biofilms using metagenomic, transcriptomic, and proteomic approaches“. In: *Frontiers in Microbiology* 8.OCT, pp. 1–14 (cit. on p. 33).

Krouk, G., B. Lacombe, A. Bielach, et al. (2010). „Nitrate-regulated auxin transport by NRT1.1 defines a mechanism for nutrient sensing in plants“. In: *Developmental Cell* (cit. on p. 27).

Kumar, A., S. N. Silim, M. Okamoto, M. Y. Siddiqi, and A. D. Glass (2003). „Differential expression of three members of the AMT1 gene family encoding putative high-affinity NH4+ transporters in roots of Oryza sativa subspecies indica“. In: *Plant, Cell and Environment* (cit. on p. 154).

Kumar, S., G. Stecher, and K. Tamura (2016). „MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets“. In: *Molecular biology and evolution* 33.7, pp. 1870–1874 (cit. on p. 96).

Kuwata, A. and M. Takahashi (1999). „Survival and recovery of resting spores and resting cells of the marine planktonic diatom Chaetoceros pseudocumisetus under fluctuating nitrate conditions“. In: *Marine Biology* (cit. on p. 4).

Laugier, E., E. Bouguyon, A. Mauries, et al. (2012). „Regulation of high-affinity nitrate uptake in roots of Arabidopsis depends predominantly on posttranscriptional control of the NRT2.1/NAR2.1 transport system“. In: *Plant Physiology* 158.2, pp. 1067–1078 (cit. on p. 155).

Lavorel, S. and E. Garnier (2002). *Predicting changes in community composition and ecosystem functioning from plant traits: Revisiting the Holy Grail* (cit. on p. 142).

Le Bescot, N., F. Mahé, S. Audic, et al. (2015). „Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding“. In: *Environmental Microbiology* 18.2, pp. 609–626 (cit. on p. 48).

Le Quéré, C., E. T. Buitenhuis, R. Moriarty, et al. (2015). „Role of zooplankton dynamics for Southern Ocean phytoplankton biomass and global biogeochemical cycles“. In: *Biogeosciences Discussions* 12.14, pp. 11935–11985 (cit. on pp. 12, 48).

Le Quere, C., S. P. Harrison, I. Colin Prentice, et al. (2005). „Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models“. In: *Global Change Biology* 11.11, pp. 2016–2040 (cit. on pp. 8, 220, 221, 235).

Lê, S., J. Josse, and F. Husson (2008). „FactoMineR : An R package for multivariate analysis“. In: *Journal of Statistical Software*. arXiv: `arXiv:0908.3817v2` (cit. on pp. 130, 158).

Lear, G., J. Bellamy, B. S. Case, J. E. Lee, and H. L. Buckley (2014). „Fine-scale spatial patterns in bacterial community composition and function within freshwater ponds“. In: *The ISME Journal* 8.8, pp. 1715–1726 (cit. on p. 124).

Leblanc, K., J. Arístegui, L. Armand, et al. (2012). „A global diatom database – abundance, biovolume and biomass in the world ocean“. In: *Earth System Science Data Discussions* (cit. on p. 22).

Leblanc, K., B. Quéguiner, F. Diaz, et al. (2018). „Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export“. In: *Nature Communications* 9.1, pp. 1–12 (cit. on p. 36).

Levitan, O., J. Dinamarca, E. Zelzion, et al. (2015). „Remodeling of intermediate metabolism in the diatom Phaeodactylum tricornutum under nitrogen stress“. In: *Proceedings of the National Academy of Sciences* 112.2, pp. 412–417 (cit. on pp. 149–151, 153, 163).

Lévy, M., O. Jahn, S. Dutkiewicz, and M. J. Follows (2014). „Phytoplankton diversity and community structure affected by oceanic dispersal and mesoscale turbulence“. In: *Limnology and Oceanography: Fluids and Environments* 4.1, pp. 67–84 (cit. on pp. 19, 79, 86).

Lévy, M., O. Jahn, S. Dutkiewicz, M. J. Follows, and F. D'Ovidio (2015). „The dynamical landscape of marine phytoplankton diversity“. In: *Journal of The Royal Society Interface* 12.111, p. 20150481 (cit. on pp. 20, 86, 126).

Li, W. K. (2002). „Macroecological patterns of phytoplankton in the northwestern North Atlantic Ocean“. In: *Nature* (cit. on p. 176).

Li, W., A. Cowley, M. Uludag, et al. (2015). „The EMBL-EBI bioinformatics web and programmatic tools framework“. In: *Nucleic Acids Research* 43.W1, W580–W584 (cit. on p. 95).

Lima-Mendez, G., K. Faust, N. Henry, et al. (2015). „Determinants of community structure in the global plankton interactome“. In: *Science*. arXiv: arXiv:1011.1669v3 (cit. on pp. 36, 232).

Litchman, E., C. a. Klausmeier, and K. Yoshiyama (2009). „Contrasting size evolution in marine and freshwater diatoms“. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.8, pp. 2665–2670 (cit. on p. 3).

Litchman, E. (2007). „Resource competition and the ecological success of phytoplankton“. In: *Evolution of Primary Producers in the Sea*, pp. 351–375 (cit. on p. 5).

Litchman, E., P. de Tezanos Pinto, K. F. Edwards, et al. (2015a). „Global biogeochemical impacts of phytoplankton: A trait-based perspective“. In: *Journal of Ecology* 103.6, pp. 1384–1396. arXiv: arXiv:1011.1669v3 (cit. on pp. 8, 91).

Litchman, E., K. F. Edwards, and C. A. Klausmeier (2015b). „Microbial resource utilization traits and trade-offs: Implications for community structure, functioning, and biogeochemical impacts at present and in the future“. In: *Frontiers in Microbiology* 6.APR, pp. 1–10 (cit. on pp. 15, 142).

Litchman, E. and C. A. Klausmeier (2008). „Trait-based community ecology of phytoplankton“. In: *Annual Review of Ecology, Evolution, and Systematics*. arXiv: 3023 (cit. on p. 11).

Litchman, E., C. A. Klausmeier, O. M. Schofield, and P. G. Falkowski (2007). „The role of functional traits and trade-offs in structuring phytoplankton communities: Scaling from cellular to ecosystem level“. In: *Ecology Letters* 10.12, pp. 1170–1181 (cit. on p. 235).

Liu, K.-H. (1999). „CHL1 is a dual-affinity nitrate transporter of Arabidopsis involved in multiple phases of nitrate uptake“. In: *The Plant Cell* (cit. on p. 27).

Liu, Y., X. Song, X. Han, and Z. Yu (2013). „Influences of external nutrient conditions on the transcript levels of a nitrate transporter gene in Skeletonema costatum“. In: *Acta Oceanologica Sinica* 32.6, pp. 82–83 (cit. on pp. 149, 150).

Logares, R., S. Audic, D. Bass, et al. (2014). „Patterns of rare and abundant marine microbial eukaryotes“. In: *Current Biology* 24.8, pp. 813–821. arXiv: 9809069v1 [arXiv:gr-qc] (cit. on p. 43).

Lomas, M. W. and P. M. Glibert (1999). „Interactions between NH + 4 and NO 3 uptake and assimilation : comparison of diatoms and dino ¯ agellates at several growth temperatures“. In: *Marin* 133, pp. 541–551 (cit. on pp. 25, 91, 150).

Lomas, M. W. and P. M. Glibert (2000). „Comparisons of nitrate uptake, stroage, and reduction in marine diatioms and flagellates“. In: *J. Phycol.* 913.3, pp. 903–913 (cit. on p. 27).

Longhi, M. L. and B. E. Beisner (2010). „Patterns in taxonomic and functional diversity of lake phytoplankton“. In: *Freshwater Biology* 55.6, pp. 1349–1366 (cit. on pp. 8, 124).

Longhurst, A., S. Sathyendranath, T. Platt, and C. Caverhill (1995). „An estimate of global primary production in the ocean from satelite radiometer data“. In: *Journal of Plankton Research* 17.6, pp. 1245–1271 (cit. on p. 17).

Luddington, I. A., I. Kaczmarska, and C. Lovejoy (2012). „Distance and character-based evaluation of the V4 region of the 18S rRNA gene for the identification of diatoms (Bacillariophyceae)“. In: *PLoS ONE* (cit. on p. 44).

Luo, C., N. Mahowald, T. Bond, et al. (2008). „Combustion iron distribution and deposition“. In: *Global Biogeochemical Cycles* (cit. on p. 198).

MacGillivary, M. L. and I. Kaczmarska (2011). „Survey of the efficacy of a short fragment of the rbcL gene as a supplemental DNA barcode for diatoms“. In: *Journal of Eukaryotic Microbiology* (cit. on p. 44).

MacIntyre, H. L., T. M. Kana, T. Anning, and R. J. Geider (2002). *Photoacclimation of photosynthesis irradiance response curves and photosynthetic pigments in microalgae and cyanobacteria* (cit. on p. 221).

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik (2016). „Cluster analysis basics and extensions. R package version 2.0.4“. In: *Cran* January 2012 (cit. on pp. 129, 158).

Magurran, A. E. (1988a). *Ecological diversity and its measurement*. Princeton University (cit. on pp. 8, 42).

Magurran, A. E. (1988b). „Why diversity?“ In: *Ecological Diversity and Its Measurement*, pp. 1–5 (cit. on p. 7).

Mahé, F., T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn (2014). „Swarm: robust and fast clustering method for amplicon-based studies“. In: *PeerJ* (cit. on p. 51).

Mallin, M. A., H. W. Paerl, and J. Rudek (1991). „Seasonal phytoplankton composition, productivity and biomass in the Neuse River estuary, North Carolina“. In: *Estuarine, Coastal and Shelf Science* (cit. on p. 5).

Malviya, S., E. Scalco, S. Audic, et al. (2016). „Insights into global diatom distribution and diversity in the world's ocean“. In: *Proceedings of the National Academy of Sciences* 113.11, E1516–E1525 (cit. on pp. 19, 36, 44, 48, 50, 82, 83, 106, 109, 125, 137, 139, 140, 144, 209, 232).

Mann, D. G. and S. J. M. Droop (1996). „Biodiversity, biogeography and conservation of diatoms“. In: *Hydrobiologia* 336.1-3, pp. 19–32 (cit. on p. 2).

Mann, D. G. (2010). „Discovering diatom species: is a long history of disagreements about species-level taxonomy now at an end?“ In: *Plant Ecology and Evolution* 143.3, pp. 251–264 (cit. on p. 44).

Mann, D. G. and P. Vanormelingen (2013). „An inordinate fondness? the number, distributions, and origins of diatom species“. In: *Journal of Eukaryotic Microbiology* (cit. on pp. 106, 231).

Marañón, E. (2015). „Cell size as a key determinant of phytoplankton metabolism and community structure“. In: *Annual Review of Marine Science* (cit. on p. 176).

Marañón, E., P. Cermeño, D. C. López-Sandoval, et al. (2013). „Unimodal size scaling of phytoplankton growth and the size dependence of nutrient uptake and use“. In: *Ecology Letters* 16.3, pp. 371–379 (cit. on pp. 168, 189, 215).

Margalef, R. (1963). „On certain unifying principles in ecology“. In: *The American Naturalist* 97.897, pp. 357–374 (cit. on p. 4).

– (1968). *Perspectives in ecological theory*. Chicago: University of Chicago press, p. 111 (cit. on p. 125).

– (1978). „Life-forms of phytoplankton as survival alternatives in an unstable environment“. In: *Oceanologica Acta* 1, pp. 493–509 (cit. on p. 4).

Marshall, J., A. Adcroft, C. Hill, L. Perelman, and C. Heisey (1997). „A finite-volume, incompressible navier stokes model for, studies of the ocean on parallel computers“. In: *Journal of Geophysical Research C: Oceans* (cit. on p. 198).

Massana, R., A. Gobet, S. Audic, et al. (2015). „Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing“. In: *Environmental microbiology* (cit. on p. 82).

Mau, B., M. a. Newton, and B. Larget (1999). „Bayesian phylogenetic inference via Markov chain Monte Carlo methods.“ In: *Biometrics* 55.1, pp. 1–12 (cit. on p. 96).

May, R. (1975). „Patterns of species abundance and diversity“. In: *Ecology and evolution of communities*. Ed. by M. Cody and J. Diamond. Cambridge: Harvard University Press, pp. 81–120 (cit. on p. 8).

McArdle, B. H. and M. J. Anderson (2001). „Fitting multivariate models to community data: A comment on distance-based redundancy analysis“. In: *Ecology* (cit. on p. 129).

Mccarthy, J. K., S. R. Smith, J. P. Mccrow, et al. (2017). „Nitrate reductase knockout uncouples nitrate transport from nitrate assimilation and drives repartitioning of carbon flux in a model pennate diatom“. In: *The Plant Cell* 29.August, pp. 2047–2070 (cit. on pp. 27, 91).

McDonald, S. M., J. N. Plant, and A. Z. Worden (2010). „The mixed lineage nature of nitrogen transport and assimilation in marine eukaryotic phytoplankton: A case study of Micromonas“. In: *Molecular Biology and Evolution* 27.10, pp. 2268–2283 (cit. on p. 102).

McGill, B. J., B. J. Enquist, E. Weiher, and M. Westoby (2006). „Rebuilding community ecology from functional traits“. In: *Trends in Ecology and Evolution* 21.4, pp. 178–185. arXiv: `arXiv:1011.1669v3` (cit. on p. 124).

McQuoid, M. R. and L. A. Hobson (1996). *Diatom resting stages* (cit. on pp. 3, 4).

Medlin, L. K. (2016). „Evolution of the diatoms: major steps in their evolution and a review of the supporting molecular and morphological evidence“. In: *Phycologia* 55.1, pp. 79–103 (cit. on pp. 102, 103).

Medlin, L. K. and I. Kaczmarska (2004). *Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision* (cit. on p. 2).

Melssen, W., R. Wehrens, and L. Buydens (2006). „Supervised Kohonen networks for classification problems“. In: *Chemometrics and Intelligent Laboratory Systems* 83.2, pp. 99–113 (cit. on p. 59).

Mende, D. R., A. S. Waller, S. Sunagawa, et al. (2012). „Assessment of metagenomic assembly using simulated next generation sequencing data“. In: *PLoS ONE* (cit. on p. 30).

Mitchell, J. G. (2018). „Whence is the diversity of diatom frustules derived?“ In: *Diatom Nanotechnology: Progress and Emerging Applications*. The Royal Society of Chemistry, pp. 1–13 (cit. on p. 3).

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, et al. (2001). „What is the observed relationship between species richness and productivity?“ In: *Ecology* 82.9, pp. 2381–2396. arXiv: 1011.1669v3 (cit. on p. 67).

Mock, T., M. P. Samanta, V. Iverson, et al. (2008). „Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses“. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.5, pp. 1579–1584. arXiv: NIHMS150003 (cit. on pp. 149, 152).

Mock, T., R. P. Otillar, J. Strauss, et al. (2017). „Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus“. In: *Nature* 541.7638, pp. 536–540 (cit. on p. 184).

Montagnes, D. J. and D. J. Franklin (2001). „Effect of temperature on diatom volume, growth rate, and carbon and nitrogen content: Reconsidering some paradigms“. In: *Limnology and Oceanography* 46.8, pp. 2008–2018 (cit. on p. 215).

Moore, J. K., S. C. Doney, and K. Lindsay (2004). *Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model* (cit. on p. 219).

Mouchet, M. A., S. Villéger, N. W. Mason, and D. Mouillot (2010). „Functional diversity measures: An overview of their redundancy and their ability to discriminate community assembly rules“. In: *Functional Ecology* 24.4, pp. 867–876 (cit. on pp. 88, 131).

Mougi, A. and M. Kondoh (2012). „Diversity of interaction types and ecological community stability“. In: *Science* (cit. on p. 10).

Muller-Karger, F. E., P. Miloslavich, N. J. Bax, et al. (2018). „Advancing marine biological observations and data requirements of the complementary Essential Ocean Variables (EOVs) and Essential Biodiversity Variables (EBVs) frameworks“. In: *Frontiers in Marine Science* 5.June, pp. 1–15 (cit. on pp. 45, 231).

Murtagh, F. and P. Legendre (2014). „Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?“ In: *Journal of Classification* 31.3, pp. 274–295. arXiv: arXiv:1111.6285v2 (cit. on pp. 128, 157).

Narasingarao, P., S. Podell, J. A. Ugalde, et al. (2012). „De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities“. In: *ISME Journal* (cit. on p. 30).

Narihiro, T. and Y. Kamagata (2013). „Cultivating yet-to-be cultivated microbes: the challenge continues“. In: *Microbes Environ.* 28, pp. 163–165 (cit. on p. 31).

Naselli-Flores, L. and J. Padisák (2016). „Blowing in the wind: how many roads can a phytoplanktont walk down? A synthesis on phytoplankton biogeography and spatial processes“. In: *Hydrobiologia* 764.1, pp. 303–313 (cit. on p. 19).

Navarro, M., R. Prieto, E. Fernandez, and A. Galvan (1996). „Constitutive expression of nitrate reductase changes the regulation of nitrate and nitrite transporters in Chlamydomonas reinhardtii“. In: *Plant J.* 9, pp. 819–827 (cit. on p. 26).

Nelson, D. M., P. Tréguer, M. A. Brzezinski, A. Leynaert, and B. Quéguiner (1995). „Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation“. In: *Global Biogeochemical Cycles* 9.3, pp. 359–372 (cit. on pp. 2, 21).

Nelson, K. E., J. L. Peterson, and S. Garges (2011). *Metagenomics of the human body*. arXiv: `arXiv:1011.1669v3` (cit. on p. 33).

Nour-Eldin, H. H., T. G. Andersen, M. Burow, et al. (2012). „NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds“. In: *Nature* (cit. on p. 27).

O'Connor, M. I., M. F. Piehler, D. M. Leech, A. Anton, and J. F. Bruno (2009). „Warming and resource availability shift food web structure and metabolism“. In: *PLoS Biology*. arXiv: `arXiv:1011.1669v3` (cit. on p. 12).

Odum, E. P. (1969). „The strategy of ecosystem development“. In: *Science* (cit. on p. 4).

Oksanen, J., F. G. Blanchet, R. Kindt, et al. (2017). *Vegan: community ecology package*. arXiv: `arXiv:1011.1669v3` (cit. on pp. 98, 128, 131, 157, 158, 199).

Paasche, E., I. Bryceson, and K. Tangen (1984). „Interspecific variation in dark nitrogen uptake by dinoflagellates“. In: *Journal of Phycology* (cit. on p. 25).

Parker, M. S. and E. V. Armbrust (2005). „Synergistic effects of light, temperature, and nitrogen source on transcription of genes for carbon and nitrogen metabolism in the centric diatom Thalassiosira pseudonana (Bacillariophyceae)“. In: *Journal of Phycology* 41.6, pp. 1142–1153 (cit. on pp. 27, 92).

Parks, M. B., T. Nakov, E. C. Ruck, N. J. Wickett, and A. J. Alverson (2018). „Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta)“. In: *American Journal of Botany* (cit. on p. 1).

Passy, S. I. and P. Legendre (2006). „Are algal communities driven toward maximum biomass?“ In: *Proceedings of the Royal Society B: Biological Sciences* 273.1601, pp. 2667–2674 (cit. on pp. 47, 66).

Passy, S. I. (2007). „Diatom ecological guilds display distinct and predictable behavior along nutrient and disturbance gradients in running waters“. In: *Aquatic Botany* 86.2, pp. 171–178 (cit. on p. 90).

Perruche, C., P. Rivière, P. Pondaven, and X. Carton (2010). „Phytoplankton competition and coexistence: Intrinsic ecosystem dynamics and impact of vertical mixing“. In: *Journal of Marine Systems* 81.1-2, pp. 99–111 (cit. on p. 19).

Pesant, S., F. Not, M. Picheral, et al. (2015). „Open science resources for the discovery and analysis of Tara Oceans data“. In: *Scientific Data* 2.Lmd, pp. 1–16 (cit. on pp. 34, 35).

Petchey, O. L. and K. J. Gaston (2006). „Functional diversity: Back to basics and looking forward“. In: *Ecology Letters* 9.6, pp. 741–758. arXiv: `arXiv:1011.1669v3` (cit. on pp. 8, 15, 89, 90).

Picheral, M., S. Searson, V. Taillandier, et al. (2014a). *Vertical profiles of environmental parameters measured from physical, optical and imaging sensors during Tara Oceans expedition 2009-2013*. data set (cit. on pp. 129, 130).

Picheral, M., S. Searson, V. Taillandier, et al. (2014b). *Vertical profiles of environmental parameters measured on discrete water samples collected with Niskin bottles during the Tara Oceans expedition 2009-2013*. data set (cit. on pp. 56, 130).

Pignatelli, M., G. Aparicio, I. Blanquer, et al. (2008). *Metagenomics reveals our incomplete knowledge of global diversity* (cit. on p. 30).

Pignatelli, M. and A. Moya (2011). „Evaluating the fidelity of De Novo short read metagenomic assembly using simulated data“. In: *PLoS ONE* (cit. on p. 30).

Pondaven, P., O. Ragueneau, P. Tréguer, et al. (2000). „Resolving the 'opal paradox' in the Southern Ocean“. In: *Nature* (cit. on p. 22).

Power, S., F. Delage, G. Wang, I. Smith, and G. Kociuba (2017). „Apparent limitations in the ability of CMIP5 climate models to simulate recent multi-decadal change in surface temperature: implications for global temperature projections“. In: *Climate Dynamics* 49.1-2, pp. 53–69 (cit. on p. 187).

Price, M. N., P. S. Dehal, and A. P. Arkin (2010). „FastTree 2 - Approximately maximum-likelihood trees for large alignments“. In: *PLoS ONE* 5.3. arXiv: `Price,MorganN.,2010,FastTree2` (cit. on pp. 51, 96).

Prince, V. E. and F. B. Pickett (2002). „Splitting pairs: The diverging fates of duplicated genes“. In: *Nature Reviews Genetics* 3.11, pp. 827–837 (cit. on p. 119).

Prowe, A. E., M. Pahlow, S. Dutkiewicz, M. Follows, and A. Oschlies (2012). „Top-down control of marine phytoplankton diversity in a global ecosystem model“. In: *Progress in Oceanography* 101.1, pp. 1–13 (cit. on pp. 10, 12).

Ptacnik, R., A. G. Solimini, T. Andersen, et al. (2008). „Diversity predicts stability and resource use efficiency in natural phytoplankton communities“. In: *Proceedings of the National Academy of Sciences* 105.13, pp. 5134–5138 (cit. on pp. 7, 42).

Rabosky, D. L., J. Chang, P. O. Title, et al. (2018). „An inverse latitudinal gradient in speciation rate for marine fishes“. In: *Nature*, p. 1 (cit. on p. 84).

Raes, J., I. Letunic, T. Yamada, L. J. Jensen, and P. Bork (2011). „Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data“. In: *Molecular Systems Biology* (cit. on p. 28).

Rappé, M. S. and S. J. Giovannoni (2003). „The uncultured microbial majority“. In: *Annual Review of Microbiology* (cit. on p. 31).

Ratnasingham, S. and P. D. N. Hebert (2013). „A DNA-Based registry for all animal species: The barcode index number (BIN) system“. In: *PLoS ONE* 8.7 (cit. on p. 68).

Raven, J. A. (1987). „The role of vacuoles“. In: *New Phytologist* (cit. on p. 5).

Rawat, S. R., S. N. Silim, H. J. Kronzucker, M. Y. Siddiqi, and A. D. Glass (1999). „AtAMT1 gene expression and NH4+ uptake in roots of Arabidopsis thaliana: Evidence for regulation by root glutamine levels“. In: *Plant Journal* (cit. on p. 154).

Rees, T., R. Cresswell, and P. Syrett (1980). „Sodium- dependent uptake of nitrate and urea by a marine diatom“. In: *Biochim. Biophys. Acta* 596, pp. 141–144 (cit. on p. 26).

Rees, T., T. R. Larson, J. Heldens, and F. Huning (1995). „In situ glutamine synthetase activity in a marine unicellular alga (development of a sensitive colorimetric assay and the effects of nitrogen status on enzyme activity).“ In: *Plant physiology* (cit. on p. 92).

Reichstein, M., M. Bahn, M. D. Mahecha, J. Kattge, and D. D. Baldocchi (2014). „Linking plant and ecosystem functional biogeography“. In: *Proceedings of the National Academy of Sciences* 111.38, pp. 13697–13702 (cit. on p. 124).

Reynolds, C. S. (1980). „Phytoplankton assemblages and their periodicity in stratifying lake systems“. In: *Ecography* (cit. on p. 89).

– (1988). „Functional morphology and the adaptive strategies of freshwater phytoplankton“. In: *Growth and Reproductive Strategies of Freshwater Phytoplankton*. Ed. by C. D. Sandgren. Cambridge: Cambridge university press, pp. 388–433 (cit. on p. 89).

Richerson, P., R. Armstrong, and C. R. Goldman (1970). „Contemporaneous disequilibrium, a new hypothesis to explain the "paradox of the plankton".“ In: *Proceedings of the National Academy of Sciences of the USA* 67.4, pp. 1710–4 (cit. on p. 11).

Ridgeway, G. (2006). „Generalized boosted regression models“. In: *Documentation on the R Package "gbm", version* 1, p. 7 (cit. on pp. 55, 159).

Riemann, L., G. F. Steward, and F. Azam (2000). „Dynamics of bacterial community composition and activity during a mesocosm diatom bloom“. In: *Applied and Environmental Microbiology* (cit. on p. 232).

Rimet, F. and A. Bouchez (2012). „Life-forms, cell-sizes and ecological guilds of diatoms in European rivers“. In: *Knowledge and Management of Aquatic Ecosystems* (cit. on p. 90).

Robinson, J. V. and C. D. Sandgren (1983). „The effect of temporal environmental heterogeneity on community structure: a replicated experimental study“. In: *Oecologia* 57.1-2, pp. 98–102 (cit. on p. 11).

Rodriguez-R, L. M. and K. T. Konstantinidis (2014). „Estimating coverage in metagenomic data sets and why it matters“. In: *ISME Journal* 8.11, pp. 2349–2351 (cit. on p. 233).

Rogato, A., A. Amato, D. Iudicone, et al. (2015). *The diatom molecular toolkit to handle nitrogen uptake* (cit. on pp. 25, 27, 92–95, 102, 104, 106, 149).

Roh, C., F. Villatte, B. Kim, and R. D. Schmid (2006). „Comparative study of methods for extraction and purification of environmental DNA from soil and sludge samples.“ In: *Applied biochemistry and biotechnology* 134.2, pp. 97–112 (cit. on p. 82).

Ronquist, F., M. Teslenko, P. van der Mark, et al. (2012). „MrBayes 3 . 2 : Efficient bayesian phylogenetic inference and model choice across a large model space“. In: *Systematic biology* 61.3, pp. 539–542 (cit. on p. 96).

Rosenfeld, J. S. (2016). „Functional redundancy in ecology and conservation“. In: *Oikos* 98.1, pp. 156–162 (cit. on p. 208).

Rosindell, J., S. P. Hubbell, and R. S. Etienne (2011). „The unified neutral theory of biodiversity and biogeography at age ten“. In: *Trends in Ecology and Evolution* 26.7, pp. 340–348 (cit. on p. 13).

Round, F. E., R. M. Crawford, and D. G. Mann (1990). *Diatoms: biology and morphology of the genera*. Cambridge university press (cit. on p. 3).

Rousseeuw, P. J. (1987). „Silhouettes: A graphical aid to the interpretation and validation of cluster analysis“. In: *Journal of Computational and Applied Mathematics* 20.C, pp. 53–65. arXiv: z0024 (cit. on pp. 129, 157).

Roy, R. S., D. C. Price, A. Schliep, et al. (2014). „Single cell genome analysis of an uncultured heterotrophic stramenopile“. In: *Scientific Reports* (cit. on p. 232).

Roy, S. and J. Chattopadhyay (2007). *Towards a resolution of 'the paradox of the plankton': A brief overview of the proposed mechanisms* (cit. on p. 10).

Rusch, D. B., A. L. Halpern, G. Sutton, et al. (2007). „The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific“. In: *PLoS Biology*. arXiv: 15334406 (cit. on p. 33).

Rynearson, T. A. and E. V. Armbrust (2005). „Maintenance of clonal diversity during a spring bloom of the centric diatom Ditylum brightwellii“. In: *Molecular Ecology* (cit. on p. 233).

Sabbe, K., K. Vanhoutte, R. L. Lowe, et al. (2001). „Six new Actinella (Bacillariophyta) species from Papua New Guinea, Australia and New Zealand: Further evidence for widespread diatom endemism in the Australasian region“. In: *European Journal of Phycology* 36.4, pp. 321–340 (cit. on p. 18).

Saitou, N. and M. Nei (1987). „The neighbor-joining method: a new method for reconstructing phylogenetic trees“. In: *Molecular Biology and Evolution* 4.4, pp. 406–425 (cit. on p. 96).

Salmaso, N., L. Naselli-Flores, and J. Padisák (2015). „Functional classifications and their application in phytoplankton ecology“. In: *Freshwater Biology* (cit. on p. 88).

Santos, A. M., F. M. Carneiro, and M. V. Cianciaruso (2015). „Predicting productivity in tropical reservoirs: The roles of phytoplankton taxonomic and functional diversity“. In: *Ecological Indicators* 48, pp. 428–435 (cit. on p. 47).

Sapriel, G., M. Quinet, M. Heijde, et al. (2009). „Genome-wide transcriptome analyses of silicon metabolism in Phaeodactylum tricornutum reveal the multilevel regulation of silicic acid transporters“. In: *PLoS ONE* 4.10 (cit. on p. 150).

Sarthou, G., K. R. Timmermans, S. Blain, and P. Tréguer (2005). *Growth physiology and fate of diatoms in the ocean: A review* (cit. on pp. 12, 215).

Scheffer, M. and E. H. van Nes (2006). „Self-organized similarity, the evolutionary emergence of groups of similar species“. In: *Proceedings of the National Academy of Sciences* (cit. on p. 13).

Schloss, P. D., S. L. Westcott, T. Ryabin, et al. (2009). „Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities“. In: *Applied and Environmental Microbiology* 75.23, pp. 7537–7541 (cit. on p. 54).

Schneider, T. D. and R. M. Stephens (1990). „Sequence logos: A new way to display consensus sequences“. In: *Nucleic Acids Research* 18.20, pp. 6097–6100 (cit. on p. 96).

Segata, N., D. Boernigen, T. L. Tickle, et al. (2013). „Computational meta'omics for microbial community studies“. In: *Molecular Systems Biology* 9.1, pp. 1–15 (cit. on p. 30).

Ser-Giacomi, E., L. Zinger, S. Malviya, C. D. Vargas, and E. Karsenti (2018). „Ubiquitous abundance distribution of non-dominant plankton across the world ' s ocean“. In: 5. arXiv: `/dx.doi.org/10.1101/269068doi: [http:]` (cit. on pp. 43, 63).

Shikata, T., M. Iseki, S. Matsunaga, et al. (2011). „Blue and red light-induced germination of resting spores in the red-tide diatom Leptocylindrus danicus“. In: *Photochemistry and Photobiology* (cit. on p. 4).

Shimoda, Y. and G. B. Arhonditsis (2016). „Phytoplankton functional type modelling: Running before we can walk? A critical evaluation of the current state of knowledge“. In: *Ecological Modelling* 320, pp. 29–43 (cit. on p. 15).

Sicko-goad, A. L. M., C. L. Schelske, E. F. Stoermer, and L. M. Sicko-goad (1984). „Estimation of intracellular carbon and silica content of diatoms from natural assemblages using morphometric techniques“. In: *Limnology and Oceanography* 29.6, pp. 1170–1178 (cit. on p. 5).

Simon, C. and R. Daniel (2011). *Metagenomic analyses: Past and future trends* (cit. on p. 28).

Sims, P. A., D. G. Mann, and L. K. Medlin (2006). „Evolution of the diatoms: insights from fossil, biological and molecular data“. In: *Phycologia* 45.4, pp. 361–402 (cit. on p. 2).

Smayda, T. (1970). „The suspension and sinking of phytoplankton in the sea“. In: *Oceanogr. Mar. Biol. Annu Rev.* 8, pp. 353–414 (cit. on p. 170).

Smetacek, V. (1999). „Diatoms and the ocean carbon cycle“. In: *Protist* 150.1, pp. 25–32 (cit. on p. 21).

Smetacek, V. S. (1985). „Role of sinking in diatom life-hystory: ecological, evolutionary and geological significance“. In: *Marine Biology* 84, pp. 239–251 (cit. on p. 6).

Smillie, C. S., M. B. Smith, J. Friedman, et al. (2011). „Ecology drives a global network of gene exchange connecting the human microbiome“. In: *Nature* (cit. on p. 33).

Smith, S. R., C. Glé, R. M. Abbriano, et al. (2016). „Transcript level coordination of carbon pathways during silicon starvation-induced lipid accumulation in the diatom Thalassiosira pseudonana“. In: *New Phytologist* 210.3, pp. 890–904 (cit. on pp. 150, 152).

Soccodato, A., F. D'Ovidio, M. Lévy, et al. (2016). „Estimating planktonic diversity through spatial dominance patterns in a model ocean“. In: *Marine Genomics* 29, pp. 9–17 (cit. on p. 79).

Sogin, M. L., M. L. Sogin, H. G. Morrison, et al. (2006). „Microbial diversity in the deep sea and the underexplored "rare biosphere".“ In: *Proceedings of the National Academy of Sciences of the United States of America*. arXiv: `arXiv:1112.4193v1` (cit. on p. 43).

Soininen, J., A. Jamoneau, J. Rosebery, and S. I. Passy (2016). „Global patterns and drivers of species and trait composition in diatoms“. In: *Global Ecology and Biogeography* 25.8, pp. 940–950 (cit. on p. 124).

Sommer, U. (1989). *Plankton ecology: succession in plankton communities*, pp. 1–369 (cit. on pp. 4, 11).

Sommer, U. (1984). *The paradox of the plankton: Fluctuations of phosphorus availability maintain diversity of phytoplankton in flow-through cultures* (cit. on pp. 5, 11).

Song, B. and B. B. Ward (2007). „Molecular cloning and characterization of high-affinity nitrate transporters in marine phytoplankton". In: *J. Phycol.* 43, pp. 542–552 (cit. on pp. 26, 91, 149, 153, 154).

Spatharis, S., D. Mouillot, D. B. Danielidis, et al. (2008). „Influence of terrestrial runoff on phytoplankton species richness-biomass relationships: A double stress hypothesis". In: *Journal of Experimental Marine Biology and Ecology* 362.1, pp. 55–62 (cit. on pp. 47, 66).

Stec, K. F., L. Caputi, P. L. Buttigieg, et al. (2017). „Modelling plankton ecosystems in the meta-omics era. Are we ready?" In: *Marine Genomics* 32, pp. 1–17 (cit. on pp. 8, 15, 28, 31, 195).

Steele, J. (1974). *The structure of marine ecosystems*. Cambridge: Harvard University Press (cit. on p. 14).

Stoeck, T., H. W. Breiner, S. Filker, et al. (2014). „A morphogenetic survey on ciliate plankton from a mountain lake pinpoints the necessity of lineage-specific barcode markers in microbial ecology". In: *Environmental Microbiology* (cit. on p. 82).

Stolte, W. and R. Riegman (1996). „A model approach for size-selective competition of marine phytoplankton for fluctuating nitrate and ammonium". In: *Journal of Phycology* (cit. on p. 5).

Strzepek, R. F. and P. J. Harrison (2004). „Photosynthetic architecture differs in coastal and oceanic diatoms". In: *Nature* (cit. on p. 4).

Suding, K. N., S. Lavorel, F. S. Chapin, et al. (2008). „Scaling environmental change through the community-level: A trait-based response-and-effect framework for plants". In: *Global Change Biology* (cit. on p. 142).

Sugihara, G., R. May, H. Ye, et al. (2012). „Detecting causality in complex ecosystems". In: *Science* (cit. on p. 177).

Sun, C., Y. Zhao, H. Li, et al. (2015). „Unreliable quantitation of species abundance based on high-throughput sequencing data of zooplankton communities". In: *Aquatic Biology* (cit. on pp. 30, 82).

Sunagawa, S., L. P. Coelho, S. Chaffron, et al. (2015). „Ocean plankton. Structure and function of the global ocean microbiome." In: *Science (New York, N.Y.)* 348.6237, p. 1261359. arXiv: NIHMS150003 (cit. on pp. 1, 124, 144, 159).

Sunda, W. G., W. G. Sunda, and S. A. Huntsman (1997). „Interrelated influence of iron , light and cell size on marine phytoplankton growth light and cell size on marine". In: *Nature* 390.November 1997, pp. 389–392 (cit. on p. 215).

Swenson, N. G. and M. D. Weiser (2010). „Plant geography upon the basis of functional traits: An example from eastern North American trees". In: *Ecology* (cit. on p. 124).

Syrett, P. (1981). „Nitrogen metabolism of microalgae". In: *Physiological Bases of Phytoplankton Ecology*. Ed. by T. Platt. Can. Bull., pp. 182–210 (cit. on p. 5).

Syrett, P. and I. Morris (1963). „The inhibition of nitrate assimilation by ammonium in chlorella“. In: *Biochimica et Biophysica Acta (BBA) - Specialized Section on Enzymological Subjects* (cit. on p. 25).

Tagliabue, A., O. Aumont, R. DeAth, et al. (2016). „How well do global ocean biogeochemistry models simulate dissolved iron distributions?“ In: *Global Biogeochemical Cycles* 30, pp. 149–174 (cit. on p. 195).

Takabayashi, M., F. P. Wilkerson, and D. Robertson (2005). „Response of glutamine synthetase gene transcription and enzyme activity to external nitrogen sources in the diatom Skeletonema costatum (Bacillariophyceae)“. In: *Journal of Phycology* 41.1, pp. 84–94 (cit. on pp. 27, 92).

Tammilehto, A., T. G. Nielsen, B. Krock, E. F. Møller, and N. Lundholm (2015). „Induction of domoic acid production in the toxic diatom Pseudo-nitzschia seriata by calanoid copepods“. In: *Aquatic Toxicology* 159, pp. 52–61 (cit. on p. 5).

Tapolczai, K., A. Bouchez, C. Stenger-Kovács, J. Padisák, and F. Rimet (2016). „Trait-based ecological classifications for benthic algae: review and perspectives“. In: *Hydrobiologia* 776.1, pp. 1–17 (cit. on p. 90).

Tautz, D. (1992). „Redundancies, development and the flow of information“. In: *Bioessays* (cit. on p. 118).

Telford, R. J., V. Vandvik, and H. J. B. Birks (2006). „Dispersal limitations matter for microbial morphospecies“. In: *Science* 312.5776, pp. 1015–1015 (cit. on p. 18).

Thomas, M., C. Kremer, C. A. Klausmeier, and E. Litchman (2012). „A global pattern of thermal adatpation in marine phytoplankton.“ In: *Science* 338.338, pp. 1085–1088 (cit. on p. 125).

Tilman, D. (1982). *Resource competition and community structure.* (Cit. on p. 11).

Tilman, D. (1977). „Resource competition between plankton algae: An experimental and theoretical approach“. In: *Ecology* 58.2, pp. 338–348. arXiv: arXiv:1011.1669v3 (cit. on p. 10).

– (2001). „Functional diversity“. In: *Encyclopedia of Biodiversity, Volume 3*, pp. 109–121 (cit. on p. 8).

Tirichine, L., A. Rastogi, and C. Bowler (2017). „Recent progress in diatom genomics and epigenomics“. In: *Current Opinion in Plant Biology* 36, pp. 46–55 (cit. on pp. 232, 233).

Török, P., E. T-Krasznai, V. B-Béres, et al. (2016). „Functional diversity supports the biomass–diversity humped-back relationship in phytoplankton assemblages“. In: *Functional Ecology* 30.9, pp. 1593–1602 (cit. on pp. 47, 88).

Totterdell, I. J., R. A. Armstrong, H. Drange, et al. (1993). „Trophic resolution“. In: *NATO ASI Series, Towards a Model of Ocean Biogeochemical Processes*. Vol. I 1O, pp. 71–92 (cit. on p. 14).

Tréguer, P. J. and C. L. De La Rocha (2013). „The world ocean silica cycle“. In: *Annual Review of Marine Science* 5.1, pp. 477–501. arXiv: arXiv:1011.1669v3 (cit. on pp. 21, 22).

Tréguer, P., C. Bowler, B. Moriceau, et al. (2018). „Influence of diatom diversity on the ocean biological carbon pump“. In: *Nature Geoscience* 11.1, pp. 27–37 (cit. on pp. 20, 125).

Tringe, S. G., C. Von Mering, A. Kobayashi, et al. (2005). „Comparative metagenomics of microbial communities“. In: *Science* (cit. on p. 30).

Turnbaugh, P., R. Ley, M. Hamady, et al. (2007). „Feature the human microbiome project“. In: *Nature*. arXiv: `arXiv:1011.1669v3` (cit. on p. 33).

Tyson, G. W., J. Chapman, P. Hugenholtz, et al. (2004). „Community structure and metabolism through reconstruction of microbial genomes from the environment“. In: *Nature*. arXiv: `arXiv:1011.1669v3` (cit. on p. 28).

Unkles, S. E., E. Karabika, V. F. Symington, et al. (2012). „Alanine scanning mutagenesis of a high-affinity nitrate transporter highlights the requirement for glycine and asparagine residues in the two nitrate signature motifs.“ In: *The Biochemical journal* 447.1, pp. 35–42 (cit. on p. 104).

Unkles, S. E., D. A. Rouch, Y. Wang, et al. (2004). „Two perfectly conserved arginine residues are required for substrate binding in a high-affinity nitrate transporter.“ In: *Proceedings of the National Academy of Sciences of the United States of America* 101.50, pp. 17549–54 (cit. on p. 104).

Valenzuela, J., A. Mazurie, R. P. Carlson, et al. (2012). „Potential role of multiple carbon fixation pathways during lipid accumulation in Phaeodactylum tricornutum“. In: *Biotechnology for Biofuels* (cit. on pp. 150, 153).

Vallina, S. M., M. J. Follows, S. Dutkiewicz, et al. (2014a). „Global relationship between phytoplankton diversity and productivity in the ocean“. In: *Nature Communications* 5, pp. 1–10 (cit. on pp. 43, 47, 62, 64, 66, 67, 125, 126, 133, 140, 141, 223, 225).

Vallina, S., B. Warda, S. Dutkiewicz, and M. Follows (2014b). „Maximal feeding with active prey-switching: A kill-the-winner functional response and its effect on global diversity and biogeography“. In: *Progress in Oceanography* 120, pp. 93–109 (cit. on pp. 12, 47, 48, 52).

Van Oostende, N., S. E. Fawcett, D. Marconi, et al. (2017). „Variation of summer phytoplanktoncommunity composition and its relationship to nitrate and regenerated nitrogen assimilation across the North Atlantic Ocean“. In: *Deep-Sea Research Part I: Oceanographic Research Papers* 121.June 2016, pp. 79–94 (cit. on p. 176).

Vandermeer, J. H. (1972). „Niche theory“. In: *Annual Review of Ecology and Systematics*. arXiv: `arXiv:1011.1669v3` (cit. on p. 11).

Vanormelingen, P., E. Verleyen, and W. Vyverman (2008). „The diversity and distribution of diatoms: From cosmopolitanism to narrow endemism“. In: *Biodiversity and Conservation* 17.2, pp. 393–405 (cit. on pp. 2, 18).

Verdy, A., M. Follows, and G. Flierl (2009). „Optimal phytoplankton cell size in an allometric model“. In: *Marine Ecology Progress Series* (cit. on p. 170).

Vergnon, R., N. K. Dulvy, and R. P. Freckleton (2009). „Niches versus neutrality: Uncovering the drivers of diversity in a species-rich community“. In: *Ecology Letters* (cit. on p. 47).

Villanova, V., A. E. Fortunato, D. Singh, et al. (2017). „Investigating mixotrophic metabolism in the model diatom phaeodactylum tricornutum“. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1728 (cit. on p. 4).

Villar, E., G. Farrant, M. Follows, et al. (2015). „Environmental characteristics of Agulhas rings affect inter-ocean plankton transport“. In: *Science* 348.6237 (cit. on p. 84).

Vincent, F. J., S. Colin, S. Romac, et al. (2018). „The epibiotic life of the cosmopolitan diatom Fragilariopsis doliolus on heterotrophic ciliates in the open ocean“. In: *ISME Journal* 12.4, pp. 1094–1108 (cit. on p. 36).

Violle, C., P. B. Reich, S. W. Pacala, B. J. Enquist, and J. Kattge (2014). „The emergence and promise of functional biogeography“. In: *Proceedings of the National Academy of Sciences* 111.38, pp. 13690–13696 (cit. on p. 20).

Violle, C., M. L. Navas, D. Vile, et al. (2007). *Let the concept of trait be functional!* (Cit. on p. 15).

Visco, J. A., L. Apothéloz-Perret-Gentil, A. Cordonier, et al. (2015). „Environmental monitoring: Inferring the diatom index from next-generation sequencing data“. In: *Environmental Science and Technology* 49.13, pp. 7597–7605 (cit. on p. 44).

Vogel, T. M., P. Simonet, J. K. Jansson, et al. (2009). „TerraGenome: A consortium for the sequencing of a soil metagenome“. In: *Nature Reviews Microbiology* (cit. on p. 33).

Volk, T. and M. I. Hoffert (1985). „Ocean carbon pumps: Analysis of relative strength and efficiencies in ocean-driven atmospheric CO_2 changes“. In: *Geophysical Monograph Series* (cit. on p. 21).

Von Dassow, P. and M. Montresor (2011). „Unveiling the mysteries of phytoplankton life cycles: Patterns and opportunities behind complexity“. In: *Journal of Plankton Research* (cit. on p. 15).

Vyverman, W., E. Verleyen, K. Sabbe, et al. (2007). „Historical processes constrain patterns in global diatom diversity“. In: *Ecology* 88.8, pp. 1924–1931 (cit. on pp. 18, 19).

Wagner, A. (2000). „Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate“. In: *Proceedings of the National Academy of Sciences* 97.12, pp. 6579–6584 (cit. on p. 155).

Wan, X. S., H. X. Sheng, M. Dai, et al. (2018). „Ambient nitrate switches the ammonium consumption pathway in the euphotic ocean“. In: *Nature Communications* 9.1, pp. 1–9 (cit. on pp. 171, 176, 177).

Wardle, D. A. (2004). „Ecological Linkages Between Aboveground and Belowground Biota“. In: *Science* 304.5677, pp. 1629–1633 (cit. on p. 8).

Waterworth, W. M. and C. M. Bray (2006). *Enigma variations for peptides and their transporters in higher plants* (cit. on p. 27).

Wehrens, R. and L. M. C. Buydens (2007). „Self- and super-organizing maps in R: The kohonen package“. In: *Journal of Statistical Software* (cit. on p. 59).

Wetzel, C. E., D. C. de Bicudo, L. Ector, et al. (2012). „Distance decay of similarity in neotropical diatom communities“. In: *PLoS ONE* 7.9, pp. 10–11 (cit. on pp. 18, 19).

Whitman, W. B., D. C. Coleman, and W. J. Wiebe (1998). „Prokaryotes: The unseen majority“. In: *Proceedings of the National Academy of Sciences* 95.12, pp. 6578–6583. arXiv: `arXiv:1011.1669v3` (cit. on p. 1).

Wilhelm, C., C. Büchel, J. Fisahn, et al. (2006). „The regulation of carbon and nutrient assimilation in diatoms is significantly different from green algae“. In: *Protist* 157.2, pp. 91–124 (cit. on p. 12).

Wilken, S., J. Huisman, S. Naus-Wiezer, and E. Van Donk (2013). „Mixotrophic organisms become more heterotrophic with rising temperature“. In: *Ecology Letters* 16.2, pp. 225–233 (cit. on p. 12).

Wilkerson, F. P., R. C. Dugdale, R. M. Kudela, and F. P. Chavez (2000). „Biomass and productivity in Monterey Bay, California: Contribution of the large phytoplankton“. In: *Deep-Sea Research Part II: Topical Studies in Oceanography* (cit. on p. 25).

Williams, D. M. (2011). „Historical biogeography, microbial endemism and the role of classification: everything is endemic“. In: *Biogeography of Microscopic Organisms. Is Everything Small Everywhere*, pp. 11–31 (cit. on p. 19).

Wittgenstein, N. J. von, C. H. Le, B. J. Hawkins, and J. Ehlting (2014). „Evolutionary classification of ammonium, nitrate, and peptide transporters in land plants“. In: *BMC Evolutionary Biology* 14.1, p. 11 (cit. on pp. 93, 102).

Worm, B. and J. E. Duffy (2003). „Biodiversity, productivity and stability in real food webs“. In: *Trends in Ecology and Evolution* 18.12, pp. 628–632 (cit. on p. 47).

Wroblewski, J. S., J. L. Sarmiento, and G. R. Flierl (1988). „An ocean basin scale model of plankton dynamics in the North Atlantic: 1. Solutions For the climatological oceanographic conditions in May“. In: *Global Biogeochemical Cycles* (cit. on p. 14).

Wunsch, C. and P. Heimbach (2007). „Practical global oceanic state estimation“. In: *Physica D: Nonlinear Phenomena* (cit. on p. 198).

Yamazaki, Y., R. Akashi, Y. Banno, et al. (2009). „NBRP databases: Databases of biological resources in Japan“. In: *Nucleic Acids Research* (cit. on p. 31).

Yang, Z. (1994). „Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods“. In: *Journal of Molecular Evolution* 39.3, pp. 306–314 (cit. on pp. 51, 96).

Zehr, J. P. and R. M. Kudela (2011). „Nitrogen cycle of the open ocean: from genes to ecosystems“. In: *Annual Review of Marine Science* (cit. on p. 23).

Zhang, J. (2003). „Evolution by gene duplication: An update“. In: *Trends in Ecology and Evolution* 18.6, pp. 292–298 (cit. on p. 155).

Zimmermann, J., G. Glöckner, R. Jahn, N. Enke, and B. Gemeinholzer (2015). „Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies“. In: *Molecular Ecology Resources* 15.3, pp. 526–542 (cit. on pp. 44, 82, 83).

Zimmermann, J., R. Jahn, and B. Gemeinholzer (2011). „Barcoding diatoms: Evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols“. In: *Organisms Diversity and Evolution* (cit. on p. 44).